

AFIT/DS/ENS/00-01

THE EVALUATION OF COMPETING CLASSIFIERS

DISSERTATION

Stephen G. Alsing  
Lieutenant Colonel, USAF

AFIT/DS/ENS/00-01

20000328 015

Approved for public release; distribution unlimited

**DTIC QUALITY INSPECTED 3**

The views expressed in this dissertation are those of the author and do not reflect the official policy or position of the Department of Defense or the United States Government.

AFIT/DS/ENS/00-01

# THE EVALUATION OF COMPETING CLASSIFIERS

DISSERTATION

Presented to the Faculty of the Graduate School of Engineering and Management  
of the Air Force Institute of Technology

Air University

In Partial Fulfillment of the  
Requirements for the Degree of  
Doctor of Philosophy in Operations Research

Stephen G. Alsing, B.S., M.S.

Lieutenant Colonel, USAF

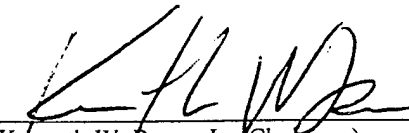
March 2000

Approved for public release; distribution unlimited

THE EVALUATION OF COMPETING CLASSIFIERS

Stephen G. Alsing, B.S., M.S.  
Lieutenant Colonel, USAF

Approved:

  
Kenneth W. Bauer, Jr. (Chairman)

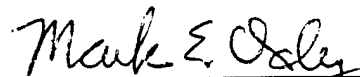
Date  
7 Mar 00

  
Meir N. Pachter (Dean's Representative)

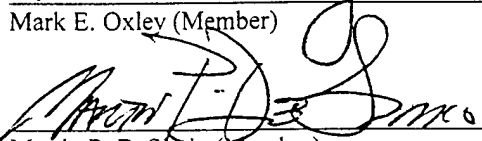
March 7, 2000

  
John O. Miller (Member)

6 Mar 00

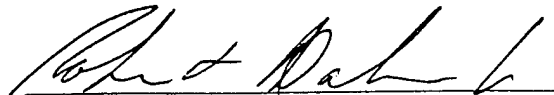
  
Mark E. Oxley (Member)

6 Mar 00

  
Martin P. DeSimio (Member)

March 6, 2000

Accepted:

  
Robert A. Calico, Jr.  
Dean, Graduate School of Engineering and Management

March 10, 2000  
Date

## *Acknowledgments*

I would not have completed this dissertation without the support of many people. The members of my committee, Lt Col J.O. Miller, Dr. Mark E. Oxley, and Dr. Martin P. Desimio provided me with support and encouragement throughout the entire process. I will always remember the many hours we spent discussing everything from mathematics to soccer, running, and life in general.

I also want to thank my Dean's Representative, Dr. Meir N. Pachter for being a great addition to my committee.

I really don't know how to thank my committee chairman, Dr Ken Bauer, enough. He got me started in my research with his "shotgun" approach and always kept me on track with innovative ideas and encouragement. I never doubted that he sincerely wanted to see me do my best and graduate. I believe he is truly motivated by his desire to teach and to help his students reach their goals. Thanks Dr Bauer.

I am thankful for the funding provided by the Air Force Research Laboratory Sensors Directorate (AFRL/SN) and the freedom they gave me to pursue my interests. I especially want to thank Capt Erik P. Blasch, Ph.D. for not only being a great AFRL/SN contact, but for being a great friend. I will always remember our times discussing research during our marathon training runs.

I thank the gang over at the centrifuge, who provided me with some adventurous and exhilarating experiences that gave me an excuse to get away from my computer now and again and see research in action.

Of course words are never enough when it comes to thanking your family. I am very grateful to my parents for providing me early on with the motivation to excel academically. My girls, Rachel and Sarah, are great. They provided me with the inspiration to keep on trying even when I thought I didn't want to.

My wife Beth means the world to me. She sure had to put up with a lot over the years. And she did. But she has always supported my dreams wherever they lead us.

Stephen G. Alsing

## *Table of Contents*

	Page
Acknowledgements . . . . .	iii
List of Figures . . . . .	ix
List of Tables . . . . .	xii
Abstract . . . . .	xvii
 I. Introduction . . . . .	 1-1
1.1 General Discussion . . . . .	1-1
1.2 Motivation . . . . .	1-1
1.2.1 ATR Problem . . . . .	1-1
1.2.2 Pilot Workload Classification Problem . . . . .	1-3
1.3 Problem Statement . . . . .	1-4
1.4 Organization of Dissertation . . . . .	1-6
 II. Literature Review . . . . .	 2-1
2.1 Overview . . . . .	2-1
2.2 Performance Assessment . . . . .	2-1
2.2.1 Confusion Matrices . . . . .	2-1
2.2.2 Error Histograms . . . . .	2-4
2.2.3 Error-Reject Curves . . . . .	2-4
2.2.4 Confidence Intervals . . . . .	2-6
2.2.5 Hypothesis Testing . . . . .	2-15
2.2.6 ROC Curves . . . . .	2-17
2.3 Performance Comparison . . . . .	2-21
2.3.1 Comparison of Confusion Matrices . . . . .	2-21
2.3.2 Comparison of Classifiers Using Hypothesis Testing . . . . .	2-24

	Page
2.3.3 Comparison of ROC Curves . . . . .	2-30
2.3.4 Multinomial Selection Procedures . . . . .	2-36
III. A Family of Metrics for Comparing Receiver Operating Characteristic Curves .	3-1
3.1 Overview . . . . .	3-1
3.2 ROC Curve Given Finite Data . . . . .	3-1
3.2.1 Mathematical Description of a ROC Curve . . . . .	3-1
3.2.2 Empirical ROC Curves from Unknown Data Distributions . .	3-3
3.3 Comparison of ROC Curves . . . . .	3-6
3.3.1 Definition of metric and metric spaces . . . . .	3-6
3.3.2 Area Under the ROC Curve (AUC) . . . . .	3-7
3.3.3 Definition of proposed ROC metrics . . . . .	3-8
3.3.4 Definition of Empirical ROC Curves . . . . .	3-10
3.3.5 ROC Convergence Theorem . . . . .	3-11
3.4 Application of Proposed ROC Metric . . . . .	3-12
3.4.1 Data Description . . . . .	3-13
3.4.2 Experiment #1 . . . . .	3-13
3.4.3 Experiment #2 . . . . .	3-15
3.5 Conclusion . . . . .	3-18
IV. Using a Multinomial Selection Procedure in Classifier Evaluation . . . . .	4-1
4.1 Overview . . . . .	4-1
4.2 Illustration of Multinomial Selection Procedure on XOR Problem . .	4-1
4.3 Block C Problem . . . . .	4-7
4.4 Iron Cross Problem . . . . .	4-10
4.5 Conclusions . . . . .	4-13

	Page
V. ATR Application . . . . .	5-1
5.1 Overview . . . . .	5-1
5.2 Data Description . . . . .	5-1
5.3 Experiment Description . . . . .	5-3
5.3.1 Experimental setup . . . . .	5-3
5.3.2 Classifiers . . . . .	5-4
5.4 Results . . . . .	5-4
5.5 Discussion . . . . .	5-7
5.6 Conclusions . . . . .	5-11
VI. Pilot Workload Application . . . . .	6-1
6.1 Overview . . . . .	6-1
6.2 Data Description . . . . .	6-1
6.3 Experiment Description . . . . .	6-3
6.3.1 Experimental setup . . . . .	6-3
6.3.2 Classifiers . . . . .	6-3
6.4 Results . . . . .	6-5
6.5 Discussion . . . . .	6-8
6.6 Conclusions . . . . .	6-9
VII. An Interpretation of Performance Measures . . . . .	7-1
7.1 Overview . . . . .	7-1
7.2 Interpretation of Classification Accuracy . . . . .	7-1
7.3 Interpretation of Area Under the ROC Curve . . . . .	7-2
7.4 Interpretation of Average Metric Distance from Diagonal . . . . .	7-3
7.4.1 2-D Normal Data Set . . . . .	7-3
7.4.2 University of Wisconsin Breast Cancer Diagnosis Data Set . . . . .	7-5
7.4.3 ATR Data Set . . . . .	7-7



	Page
7.4.4 Pilot Workload Data Set . . . . .	7-9
7.5 Interpretation of Probability of Being the Best . . . . .	7-11
7.5.1 2-D Normal Data Set . . . . .	7-12
7.5.2 University of Wisconsin Breast Cancer Diagnosis Data Set . .	7-12
7.6 Conclusions . . . . .	7-16
VIII. Summary and Recommendations . . . . .	8-1
8.1 Overview . . . . .	8-1
8.2 Contributions . . . . .	8-1
8.2.1 Development of the Signal-to-Noise Ratio (SNR) Screening Method	8-1
8.2.2 Background Reference on Performance Assessment and Performance Comparison Methods Used in Classifier Evaluation . .	8-1
8.2.3 Proof of Convergence of Receiver Operating Characteristic (ROC) Curves . . . . .	8-1
8.2.4 Development of a New Methodology for Comparing ROC Curves	8-2
8.2.5 Development of a New Methodology Using a Multinomial Selection Procedure for Comparing Competing Classifiers . . .	8-2
8.3 Recommendations for Future Research . . . . .	8-2
8.3.1 Application of New Methodologies for Evaluating Competing Classifiers to Other Classifier Types and Other Problems . .	8-2
8.3.2 Extension of New Methodology for Comparing ROC Curves to Multiple Probability Measures . . . . .	8-3
8.3.3 Development of a Systematic Methodology for Using Both Typical Performance Measures and Proposed Measures . . . . .	8-3
8.3.4 Development of a Hybrid Classifier Using the Classification Results of Competing Classifiers . . . . .	8-3
Appendix A. Proof of ROC Convergence Theorem . . . . .	A-1
A.1 Overview . . . . .	A-1

	Page
A.2 Proof that the probabilities of false positive and true positive are consistent estimators . . . . .	A-1
A.3 Proof of pointwise convergence for the estimated probability pair . . .	A-3
A.4 Proof that the integral of a real-valued random variable converges . .	A-4
A.5 Proof of the convergence of the sequence of ROC curves . . . . .	A-11
Appendix B. Glossary of Acronyms and Abbreviations . . . . .	B-1
Bibliography . . . . .	BIB-1
Vita . . . . .	VITA-1

## List of Figures

Figure		Page
1.1.	The three major components of MSTAR. . . . .	1-2
1.2.	Feedforward multilayer perceptron ANN. The inputs $x_i$ represent the features used for classification while $z_k$ are the outputs generated by the ANN to determine the classification. The hidden nodes $y_j$ , the bias, and the connection weights $w_{i,j}^{1,2}$ are the parameters used by the ANN. . . . .	1-4
1.3.	Example of typical classifier comparison using notional cancer detection problem. . . . .	1-5
2.1.	Typical gray level confusion matrices. (a) shows a good CS with few classification errors, (b) shows a poor CS with many classification errors, and (c) shows a CS with structured errors. . . . .	2-3
2.2.	Error Histogram example for a well trained classifier . . . . .	2-5
2.3.	Error-Reject tradeoff curve. . . . .	2-6
2.4.	Venn diagram showing the relationship between modeled, training, and testing conditions [61]. . . . .	2-12
2.5.	Venn diagram showing the relationship between accuracy, robustness, extensibility, and utility [61]. . . . .	2-13
2.6.	Typical ROC curve. . . . .	2-17
2.7.	Target and clutter pdfs for a two class problem. Target: $N(0, 1.5)$ ; Clutter: $N(4, 2)$ . . . . .	2-18
2.8.	Relationships and terms associated with ROC analysis (adapted from Hildebolt <i>et al.</i> , 1991 [40]). . . . .	2-20
2.9.	Generation of ROC curve (right) using two 1-D normal distributions (left). Target: $N(0, 1.5)$ ; Clutter: $N(4, 2)$ . . . . .	2-21
2.10.	Qualitative comparison of ROC curves. Competitor A is <i>better</i> than Competitor B. . . . .	2-22
2.11.	Effective range of areas ( $A$ ) under ROC curves. . . . .	2-31
2.12.	ROC curve generated from 109 CT images using the discrete rating system (data from McNeil and Hanley, 1982 [49]). . . . .	2-33

Figure		Page
2.13.	ROC curve plotted using normal-deviate coordinates. . . . .	2-34
3.1.	Both $f$ and $g$ have exactly the same area, but they are clearly not the same function. . . . .	3-9
3.2.	Pictorial representation of the convergence in the Hausdorff metric of the set $\mathcal{D}^{(n)}$ for $n$ data points to the set $S$ of all data points. . . . .	3-11
3.3.	Experiment 1: average ROC curves for three MLP Classifiers. ROC Curves averaged over 30 different test data sets. . . . .	3-14
3.4.	Experiment 2: average ROC Curves for linear, quadratic, and MLP classifiers. ROC curves averaged over 30 different test data sets. . . . .	3-16
4.1.	The exclusive-OR problem, also known as XOR, consists of data belonging to one of two classes $C_1$ and $C_2$ which are not linearly separable. . . . .	4-2
4.2.	ROC curves for XOR problem. . . . .	4-3
4.3.	The Block C problem consists of data belonging to one of two classes $C_1$ and $C_2$ which are not linearly separable (taken from Belue, 1995 [17]). . . . .	4-10
4.4.	ROC curves for Block C data. . . . .	4-11
4.5.	A very challenging discrimination problem, termed the Iron Cross problem. .	4-13
4.6.	ROC curves for Iron Cross data. . . . .	4-14
5.1.	Average HRR profile for a BMP2 (serial # C21) armored personnel carrier. Depression angle is $17^\circ$ and aspect angle is $150.1914^\circ$ . . . . .	5-3
5.2.	Average ROC curves for ATR application. Curves averaged over 33 independent test data sets. . . . .	5-6
6.1.	Plot of classification accuracy (on independent test data set) vs. number of features generated by the Signal-to-Noise Ratio (SNR) algorithm [14] for the pilot workload application. . . . .	6-4
6.2.	Average ROC curves for Pilot Workload application. Curves averaged over 30 different test data sets. . . . .	6-7
7.1.	ROC curves for unperturbed 2-D Normal data. . . . .	7-5

Figure		Page
7.2.	Classification accuracy as a function of perturbation fraction for linear, quadratic, and MLP classifiers for 2-D normal data set. . . . .	7-6
7.3.	Average ROC curves for University of Wisconsin Breast Cancer Diagnosis Data Set. ROC curves averaged over 30 different test data sets. . . . .	7-7
7.4.	Mean classification accuracy as a function of perturbation fraction for linear, quadratic, and MLP classifiers for University of Wisconsin Breast Cancer Diagnosis Data Set. . . . .	7-8
7.5.	ROC curves for one particular ATR test data set. . . . .	7-9
7.6.	Classification accuracy as a function of perturbation fraction for linear, quadratic, and MLP classifiers for one particular ATR test data set. . . . .	7-10
7.7.	ROC curves for one particular pilot workload test data set. . . . .	7-10
7.8.	Classification accuracy as a function of perturbation fraction for linear, quadratic, and MLP classifiers for one particular pilot workload test data set. . . . .	7-11
7.9.	Estimated Linear posterior probability contour plot for 2-D Normal Data Set. (Dots represent actual Class 2 data points and X's mark class boundary). . .	7-13
7.10.	Estimated Quadratic posterior probability contour plot for 2-D Normal Data Set. (Dots represent actual Class 2 data points and X's mark class boundary). . .	7-14
7.11.	Estimated MLP posterior probability contour plot for 2-D Normal Data Set. (Dots represent actual Class 2 data points and X's mark class boundary). . .	7-15
7.12.	Average estimated Linear posterior probability contour plot for University of Wisconsin Breast Cancer Diagnosis Data Set (Dots represent actual Class 2 data points and X's mark class boundary). . . . .	7-18
7.13.	Average estimated Quadratic posterior probability contour plot for University of Wisconsin Breast Cancer Diagnosis Data Set (Dots represent actual Class 2 data points and X's mark class boundary). . . . .	7-19
7.14.	Average estimated MLP posterior probability contour plot for University of Wisconsin Breast Cancer Diagnosis Data Set (Dots represent actual Class 2 data points and X's mark class boundary). . . . .	7-20

## *List of Tables*

Table		Page
1.1.	Technical barriers facing performance estimation for MSTAR. The barriers in italics represent the original motivation for this research. . . . .	1-2
1.2.	Confusion matrices for notional cancer detection problem. . . . .	1-5
2.1.	Example Confusion Matrix. . . . .	2-2
2.2.	Example Population Confusion Matrix. . . . .	2-2
2.3.	Example Composite Confusion Matrix. . . . .	2-2
2.4.	MSTAR EOC Class and Target Types . . . . .	2-7
2.5.	Example of SAR ATR dimensions by group . . . . .	2-11
2.6.	Random assignment confusion matrix. . . . .	2-22
2.7.	A priori assignment confusion matrix. . . . .	2-22
2.8.	Values of $u$ for the full range of possible $p_1$ and $p_2$ values [21]. . . . .	2-28
2.9.	Values of $u$ for selected $p_1$ and $p_2$ values from 0.66 to 0.75 [21]. . . . .	2-28
2.10.	Typical discrete rating system used in medical studies. . . . .	2-32
2.11.	Discrete rating of 109 CT images (data from McNeil and Hanley, 1982 [49]). . . . .	2-32
2.12.	$2 \times 2$ Confusion matrices for varying decision thresholds applied to 109 CT images (data from McNeil and Hanley, 1982 [49]). . . . .	2-33
3.1.	Feature rankings for University of Wisconsin Breast Cancer Diagnosis Data Set obtained by using the signal-to-noise ratio algorithm [14]. Rankings averaged over 30 different test data sets. . . . .	3-13
3.2.	Comparison of area under ROC curves and average metric distances from diagonal line for MLP ROCs. Bonferroni confidence intervals based on 30 different test data sets (normal assumption with $\alpha_{total} = 0.05$ ). . . . .	3-17
3.3.	Distance matrix showing absolute area differences (lower triangular matrix) and average metric distances (upper triangular matrix) between MLP ROC curves. Bonferroni confidence intervals based on 30 different test data sets (normal assumption with $\alpha_{total} = 0.05$ ). . . . .	3-17

Table		Page
3.4.	Comparison of area under ROC curves and average metric distances from diagonal line for linear, quadratic, and MLP classifiers. Bonferroni confidence intervals based on 30 different test data sets (normal assumption with $\alpha_{total} = 0.05$ ). . . . .	3-17
3.5.	Distance matrix showing absolute area differences (lower triangular matrix) and average metric distances (upper triangular matrix) between linear, quadratic, and MLP classifiers. Bonferroni confidence intervals based on 30 different test data sets (normal assumption with $\alpha_{total} = 0.05$ ). . . . .	3-17
4.1.	Linear classifier confusion matrix for XOR data. . . . .	4-3
4.2.	Quadratic classifier confusion matrix for XOR data. . . . .	4-3
4.3.	MLP classifier confusion matrix for XOR data. . . . .	4-3
4.4.	AUCs (area under ROC curve) for XOR data. . . . .	4-5
4.5.	BEM procedure illustrated for Class 1 XOR data. . . . .	4-5
4.6.	BEM procedure illustrated for Class 2 XOR data. . . . .	4-5
4.7.	Estimates and Bonferroni confidence intervals ( $\alpha_{total} = 0.05$ ) for the probability of being the best classifier given Class 1 data . . . . .	4-6
4.8.	Estimates and Bonferroni confidence intervals ( $\alpha_{total} = 0.05$ ) for the probability of being the best classifier given Class 2 data . . . . .	4-6
4.9.	Estimates and Bonferroni confidence intervals ( $\alpha_{total} = 0.05$ ) for the total probability of being the best classifier for XOR data. . . . .	4-6
4.10.	Values for the Probability of Correct Selection ( <i>PCS</i> ) for various number of test data points for the XOR problem. . . . .	4-8
4.11.	Linear classifier confusion matrix for Block C data. . . . .	4-9
4.12.	Quadratic classifier confusion matrix for Block C data. . . . .	4-9
4.13.	MLP classifier confusion matrix for Block C data. . . . .	4-9
4.14.	AUCs (area under ROC curve) for Block C data. . . . .	4-9
4.15.	Estimates and Bonferroni confidence intervals ( $\alpha_{total} = 0.05$ ) for the total probability of being the best classifier for Block C data. . . . .	4-10
4.16.	Linear classifier confusion matrix for Iron Cross data. . . . .	4-12

Table		Page
4.17.	Quadratic classifier confusion matrix for Iron Cross data. . . . .	4-12
4.18.	MLP classifier confusion matrix for Iron Cross data. . . . .	4-12
4.19.	AUCs (area under ROC curve) for Iron Cross data. . . . .	4-12
4.20.	Estimates and Bonferroni confidence intervals ( $\alpha_{total} = 0.05$ ) for the total probability of being the best classifier for Iron Cross data. . . . .	4-13
5.1.	Target listing for MSTAR Public Data Set. . . . .	5-2
5.2.	Linear classifier confusion matrix for ATR application. Raw numbers are the sums over 33 test data sets. Percentages and corresponding half-lengths of the Bonferroni confidence intervals are also based on 33 independent test data sets (normal assumption with $\alpha_{total} = 0.05$ ). . . . .	5-5
5.3.	Quadratic classifier confusion matrix for ATR application. Raw numbers are the sums over 33 test data sets. Percentages and corresponding half-lengths of the Bonferroni confidence intervals are also based on 33 independent test data sets (normal assumption with $\alpha_{total} = 0.05$ ). . . . .	5-5
5.4.	MLP classifier confusion matrix for ATR application. Raw numbers are the sums over 33 test data sets. Percentages and corresponding half-lengths of the confidence intervals are also based on 33 independent test data sets (normal assumption with $\alpha = 0.05$ ). . . . .	5-5
5.5.	Summary of various performance measures for ATR application. Mean estimates and half-lengths for simultaneous Bonferroni confidence intervals based on 33 independent test data sets (normal assumption with $\alpha_{total} = 0.05$ ). . .	5-7
5.6.	Comparison of the trapezoidal approximation (AUC), the binormal approximation ( $A_z$ ), and the Wilcoxon approximation ( $W$ ) for computing the area under the ROC curve for each classifier for the ATR application. Mean estimates and half-lengths for simultaneous Bonferroni confidence intervals based on 33 independent test data sets (normal assumption with $\alpha_{total} = 0.05$ ). . .	5-8
5.7.	Comparison of the classifiers using the modified metric distance. . . . .	5-9
5.8.	Estimates for the probability of being the best classifier given the target class for the ATR application. Mean estimates and half-lengths for Bonferroni confidence intervals based on 30 different test data sets. . . . .	5-10



Table	Page
5.9. Updated estimates and Bonferroni confidence intervals ( $\alpha_{total} = 0.05$ ) for the total and individual class probabilities of the quadratic and MLP classifiers being the best for the ATR application after the linear classifier is removed from consideration. . . . .	5-10
6.1. Listing of 14 most salient features identified by SNR algorithm for pilot workload application. . . . .	6-4
6.2. Linear classifier confusion matrix for Pilot Workload application. Raw numbers are the sums over 30 test data sets. Percentages and corresponding half-lengths of the confidence intervals (CIs) are also based on 30 different test data sets (Bonferroni CIs using normal assumption with $\alpha_{total} = 0.05$ ). . . . .	6-6
6.3. Quadratic classifier confusion matrix for Pilot Workload application. Raw numbers are the sums over 30 test data sets. Percentages and corresponding half-lengths of the confidence intervals are also based on 30 different test data sets (Bonferroni CIs using normal assumption with $\alpha_{total} = 0.05$ ). . . . .	6-6
6.4. MLP classifier confusion matrix for Pilot Workload application. Raw numbers are the sums over 30 test data sets. Percentages and corresponding half-lengths of the confidence intervals are also based on 30 different test data sets (Bonferroni CIs using normal assumption with $\alpha_{total} = 0.05$ ). . . . .	6-6
6.5. Summary of various performance measures for Pilot Workload application. Mean estimates and half-lengths for Bonferroni confidence intervals (normal assumption with $\alpha_{total} = 0.05$ ) based on 30 different test data sets. . . . .	6-8
6.6. Estimates for the probability of being the best classifier given the workload class for Pilot Workload application. Mean estimates and half-lengths for Bonferroni confidence intervals based on 30 different test data sets. . . . .	6-9
7.1. Example Confusion Matrix for computing classification accuracy. . . . .	7-1
7.2. Classification accuracy and average metric distance from diagonal for unperturbed 2-D normal data. . . . .	7-4
7.3. Mean classification accuracies and average metric distances from diagonal line for linear, quadratic, and MLP classifiers. Bonferroni confidence intervals based on 30 different test data sets (normal assumption with $\alpha_{total} = 0.05$ ). . . . .	7-6

Table		Page
7.4.	Classification accuracy and average metric distance from diagonal for one ATR test data set. . . . .	7-8
7.5.	Classification accuracy and average metric distance from diagonal for one pilot workload test data set. . . . .	7-10
7.6.	Estimates and Bonferroni confidence intervals ( $\alpha_{total} = 0.05$ ) for the probability of being the best classifier given Class 2 data for 2-D Normal Data Set. . . . .	7-12
7.7.	Means and Bonferroni confidence intervals ( $\alpha_{total} = 0.05$ ) for the probability of being the best classifier given Class 2 data for University of Wisconsin Breast Cancer Diagnosis Data Set. . . . .	7-16

*Abstract*

This dissertation research makes contributions towards the objective evaluation of competing classifiers, i.e., classification systems (CSs) or pattern recognition algorithms. Automatic CSs have been under development for almost 40 years in a wide range of military and medical applications. During this period, scientists and engineers have developed extensive theory and algorithms for classification, but by comparison have focused little on the testing and evaluation of their systems. Classifier evaluation is very important in the fields of automatic target recognition (ATR) and pilot workload classification. In order for military operators to be confident in new CSs, they must have an objective way of testing and evaluating competing systems.

The purpose of this dissertation research is to advance the knowledge of classifier evaluation. The basis of the research is a commonly used evaluation tool in ATR and medical applications, called the receiver operating characteristic (ROC) curve. A proof of convergence with respect to increasing sample size for these ROC curves is provided. This ROC convergence theorem is important because it provides the basis for a framework for the comparison of ROC curves and hence, the comparison of classifiers. A demonstration is given to show how this framework can be employed using metrics that provide more insight about classifier differences than the typical area under the curve performance index used in ROC analysis. As an alternative to ROC type analyses, a method for using a multinomial selection procedure to evaluate competing classifiers is presented and demonstrated. A comparison is then made between the methodologies introduced in this research and typical approaches. Both ATR and pilot workload applications are used to make these comparisons. A review of the interpretations of the typical performance measures used is given along with interpretations for the proposed performance measures introduced in this dissertation. Finally, research contributions are summarized and future directions highlighted.

# THE EVALUATION OF COMPETING CLASSIFIERS

## *I. Introduction*

### *1.1 General Discussion*

This dissertation research makes contributions towards the objective evaluation of competing classifiers, i.e., classification systems (CSs) or pattern recognition algorithms. Automatic CSs have been under development for almost 40 years in a wide range of military and medical applications. During this period, scientists and engineers have developed extensive theory and algorithms for classification, but by comparison have focused little on the testing and evaluation of their systems. The issue of classifier evaluation is very important in the fields of automatic target recognition (ATR) and pilot workload classification where data are finite. In order for military operators to be confident in new CSs, they must have an objective way of testing and evaluating competing systems.

### *1.2 Motivation*

*1.2.1 ATR Problem.* The United States Air Force (USAF) is especially interested in objectively evaluating algorithm upgrades to their ATR system named MSTAR (Moving and Stationary Target Acquisition and Recognition) [2]. The MSTAR system is a model-based approach to automatic target recognition of synthetic aperture radar (SAR) imagery. Previous approaches to the SAR ATR problem relied on vast data libraries of targets at numerous aspect and depression angles as well as different configurations (e.g., tank hatch open/closed). The model-based approach relies on computer generated templates for matching a specific identity to each image, using only a small data library of actual stored SAR images of targets [21]. The MSTAR system consists of three major components shown in Figure 1.1. These components are [41, 42]:

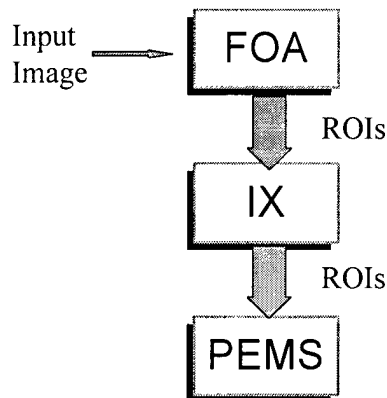


Figure 1.1 The three major components of MSTAR.

Table 1.1 Technical barriers facing performance estimation for MSTAR. The barriers in *italics* represent the original motivation for this research.

1	<i>Undersampling of mission space</i>
2	<i>ATR Performance vs. Unknowns</i>
3	<i>False Alarm Performance</i>
4	Relationship Between Mission and Extended Operating Conditions (EOC) Parameters
5	Synthetic Data
6	Data Truthing
7	<i>Performance Theory</i>
8	Joint Human/ATR Performance
9	Modeling & Simulation for ATR Evaluation

1. Focus of Attention (FOA) module identifies regions of interest (ROIs) in the image.
2. Index (IX) module generates a list of hypotheses (target/orientation) for a ROI.
3. Predict/Extract/Match/Search (PEMS) loop performs final classification of a ROI.

A change in one of these components constitutes a new MSTAR configuration or upgrade which must be evaluated objectively for its performance [21]. The Air Force Research Laboratory Sensors Directorate (AFRL/SN) at Wright-Patterson Air Force Base, Ohio manages the MSTAR program and is directing its research efforts toward investigating solutions to the technical barriers (Table 1.1) facing performance estimation for MSTAR [3]. These technical barriers, specifically performance theory, the undersampling of mission space, ATR performance vs. unknowns, and

false alarm performance are the original motivation for this research. However, this research is directly applicable to a variety of applications, including the classification of pilot workload which is another high priority research area of the USAF.

*1.2.2 Pilot Workload Classification Problem.* The issue of pilot workload is important to the USAF because pilot overload or task saturation can decrease mission effectiveness and, in some extreme cases, cause loss of life [9]. The modern aircraft, especially the military fighter is not an ideal work station for human operators. The fighter pilot must perform complex cognitive tasks while experiencing acceleration levels up to +9 Gs [34]. Between 1986 and 1995, the USAF lost 14 fighter pilots to G-induced loss of consciousness. All but one of these 14 mishaps occurred during high workload, demanding portions of the flight. These mishaps resulted because the pilots were overly task saturated and therefore unable to perform adequate anti-G straining maneuvers [9]. The ultimate goal of pilot workload research is to put instrumentation in every cockpit to monitor a pilot's workload and to warn a pilot that overload or task saturation is imminent [34].

Previous research to classify pilot workload has used psychophysiological measures such as heart rate, heart rate variability, respiration rate, respiration rate variability, and eye blink rate [33]. Measures of on-going brain electrical activity, as measured by electroencephalograph (EEG), have only been recently added to the arsenal of pilot workload measurements [33]. Artificial neural networks (ANNs) have shown great promise for classifying pilot workload using both EEG and psychophysiological measures [33–35]. ANNs have been successful because of the nonlinearity of the workload data and the generalization capabilities of ANNs [1, 33]. A significant amount of previous research to classify pilot workload has used ANNs and, in particular, feedforward multilayer perceptron (MLP) ANNs. A typical feedforward MLP ANN is shown in Figure 1.2.

The inputs to these feedforward ANNs typically include peripheral psychophysiological features as well as features preprocessed in a variety of ways from EEG. Unfortunately, irrelevant input features to an ANN can reduce classifier performance. In order to identify the important

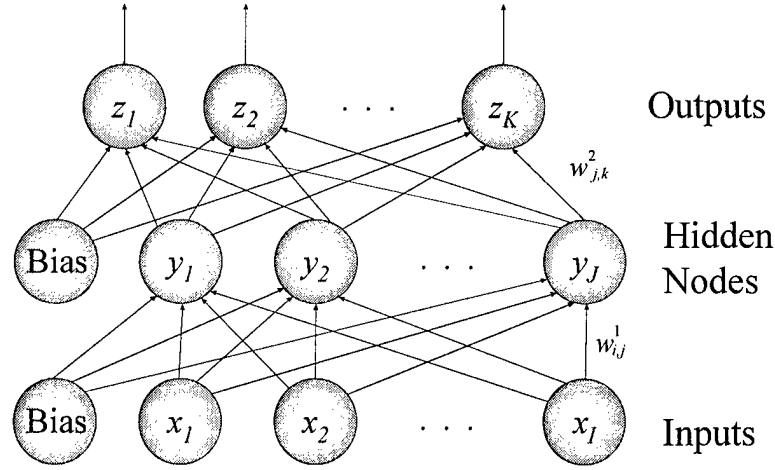


Figure 1.2 Feedforward multilayer perceptron ANN. The inputs  $x_i$  represent the features used for classification while  $z_k$  are the outputs generated by the ANN to determine the classification. The hidden nodes  $y_j$ , the bias, and the connection weights  $w_{i,j}^{1,2}$  are the parameters used by the ANN.

input features in a MLP ANN with many input features, the initial stages of this research resulted in the development of the Signal-to-Noise Ratio (SNR) saliency measure and screening method for selecting a parsimonious set of features [14]. Greene et al. [33–35] successfully applied this SNR screening method to determine which EEG and psychophysiological features are relevant for classifying mental workload via a feedforward ANN.

In all of the research on the mathematical modeling of pilot workload, classification accuracy has been used as the sole performance measure to compare different models. Other evaluation tools are available. As part of this research, these tools are reviewed and examined in order to develop mathematically rigorous selection procedures to evaluate competing models of pilot workload.

### 1.3 Problem Statement

One of the problems facing the pattern recognition community is the question of how to objectively evaluate competing classifiers. In many applications only one performance measure, typically classification accuracy (CA) is used to distinguish between competing classifiers. For example, consider the classification results for a notional cancer detection problem shown in Figure

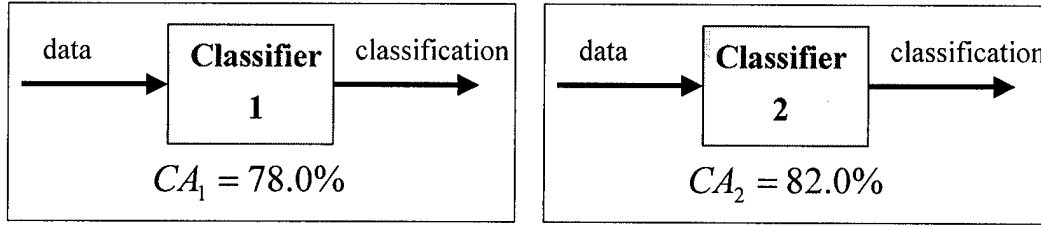


Figure 1.3 Example of typical classifier comparison using notional cancer detection problem.

Table 1.2 Confusion matrices for notional cancer detection problem.

Classifier 1 ( $CA = 78.0\%$ )

CLASSIFICATION			
T R U T H		Normal	Cancer
	Normal	1616 (76.2%)	504 (23.8%)
	Cancer	65 (14.0%)	400 (86.0%)

$$P_{FA} = 23.8\%$$

$$P_D = 86.0\%$$

Classifier 2 ( $CA = 82.0\%$ )

CLASSIFICATION			
T R U T H		Normal	Cancer
	Normal	2120 (100%)	0 (0.0%)
	Cancer	465 (100%)	0 (0.0%)

$$P_{FA} = 0\%$$

$$P_D = 0\%$$

1.3. Classifier 2 has  $CA = 82.0\%$  which means that Classifier 2 was successful 82% of the time identifying both cancer and non-cancer images alike. If Classifier 2's  $CA = 82.0\%$  is significantly greater (both statistically and practically) than Classifier 1's  $CA = 78.0\%$ , then Classifier 2 would be considered the better classifier.

Depending on the particular problem, one performance measure may not always be sufficient. Consider the classification results for the notional cancer problem displayed in more detail as confusion matrices in Table 1.2. Confusion matrices (Section 2.2.1) show classification results vertically down the columns compared to truth which is shown horizontally across the rows. Classifier 2 classified all the normal images as normal, achieving a probability of false alarm  $P_{FA}$  equal to 0%. However, Classifier 2 failed to classify any of the cancer images as cancer, achieving a probability of detection  $P_D$  equal to 0%. While Classifier 1 failed to classify all the normal images as normal, achieving  $P_{FA} = 23.8\%$ , Classifier 1 did identify a good percentage of cancer images as cancer, achieving  $P_D = 86.0\%$ . Classifier 1 may be the better classifier if the goal is to identify cancer images correctly rather than achieving an overall high CA. However, even these results, which



provide more information that classification accuracy alone, do not tell the whole story. The results shown in Table 1.2 depend upon a particular decision threshold for declaring a cancer image as cancer. A receiver operating characteristic (ROC) curve (Figure 2.6, page 2-17 and Sections 2.2.6 and 2.3.3) shows the relationship between  $P_{FA}$  and  $P_D$  as the decision threshold is varied from a very conservative value, i.e., a value that results in zero probability of detection and zero probability of false alarm, to a very aggressive value, i.e., a value that results in 100% probability of detection and 100% probability of false alarm.

ROC curves are commonly used as an evaluation tool in ATR and medical applications. An implicit assumption in the literature is that for the case of unlimited data, a limiting ROC curve exists. The major thrust of this research is the introduction of a family of metrics for comparing ROC curves that enable a proof of convergence for these curves, while also providing a useful tool for distinguishing between competing classifiers. As an alternative to ROC type analyses, a method for using a multinomial selection procedure to evaluate competing classifiers is also explored. These two methods represent differing world views of the classifier comparison problem. The methods are compared and contrasted on real world problems.

#### *1.4 Organization of Dissertation*

The remainder of this dissertation is organized as follows. Chapter II provides a literature review of performance assessment and performance comparison of CSs. Chapter III introduces a family of metrics for comparing ROC curves that enable a proof of convergence for these curves. This ROC convergence theorem is important because it provides the basis for a framework for the comparison of ROC curves and hence, the comparison of classifiers. A demonstration is also provided in this chapter to show how this framework can be employed using metrics that provide more insight about classifier differences than the typical area under the curve performance index used in ROC analysis. Chapter IV introduces a multinomial selection procedure as an alternative

to ROC type analyses for evaluating competing classifiers. Chapter V and VI provide comparisons between the methodologies introduced in this dissertation and typical approaches on real-world problems. Chapter V summarizes the results obtained using various methodologies for comparing competing classifiers for an ATR application using the MSTAR public release data set. Chapter VI summarizes the results obtained using various methodologies for a pilot workload classification problem. Chapter VII provides interpretations of the typical performance measures used in comparing competing classifiers as well as interpretations for the new performance measures introduced in this dissertation. Research contributions are summarized and future directions highlighted in Chapter VIII. Appendix A contains the proof of the ROC convergence theorem and Appendix B provides a glossary of acronyms and abbreviations. This research has resulted in many publications [1–7, 14, 19].

## *II. Literature Review*

### *2.1 Overview*

This chapter reviews the pertinent literature on the two main topic areas required to complete this dissertation research—performance assessment and performance comparison of classification systems (CSs). The majority of the following discussion is a summary of a technical report entitled, “Survey of Statistical Analysis and Experimental Design in ATR Evaluation” [2]. Therefore, the literature review presented here has a definite ATR slant. However, the performance assessment and performance comparison methods described in this chapter apply equally as well to a wide variety of other classification and detection problems.

This chapter is organized into two main sections. The performance assessment section contains a review of typical classifier performance assessment techniques, which include the use of confusion matrices, error-reject curves, confidence intervals, hypothesis testing, and receiver operating characteristic (ROC) curves. The section on performance comparison begins by describing the comparison of confusion matrices for competing classifiers. This section also discusses the comparison of classifiers using non-sequential and sequential hypothesis testing. Special attention is given in this section to the discussion of the comparison of different ROC curves representing different classifiers. Finally, the last part of this section presents an overview of multinomial selection procedures.

### *2.2 Performance Assessment*

*2.2.1 Confusion Matrices.* The easiest way to report the classification results of a CS is through the use of a discrimination event matrix (term used by ATR community [11]) or more commonly referred to as a confusion matrix in the pattern recognition community [24]. The confusion matrix is a square matrix with a single row and single column for each category defined in the data set. The rows of the matrix relate to the actual (ground truth) membership while the columns give

Table 2.1 Example Confusion Matrix.  
Classified As (Reported)

Actual (Truth)		Target 1	Target 2	Target 3	Non-Target
	Target 1	24	0	1	5
	Target 2	1	25	1	3
	Target 3	2	3	20	5
	Non-Target	1	3	4	82

Table 2.2 Example Population Confusion Matrix.  
Classified As (Reported)

Actual (Truth)		Target 1	Target 2	Target 3	Non-Target
	Target 1	80.0%	0.0%	3.3%	16.7%
	Target 2	3.3%	83.3%	3.3%	10.0%
	Target 3	6.7%	10.0%	66.7%	16.7%
	Non-Target	1.1%	3.3%	4.4%	91.1%

the predicted (CS reported) membership. Table 2.1 illustrates the confusion matrix format for a notional ATR example. The  $(i, j)$  entry in the matrix is the number of ATR reports on target  $j$  (predicted classifications) that correspond to ground truth target  $i$  (actual class membership). For example, the  $(3, 1)$  entry of the matrix indicates that the ATR reported two target 1 types which were actually target 3 types. A perfect ATR system for this example would have  $(30, 30, 30, 90)$  along the diagonal and zeros elsewhere. Rather than using the raw numbers in the confusion matrix as in Table 2.1, some CS designers will report the population counterpart (conditioned on the rows) with entries that are percentages as indicated in Table 2.2. Another reporting alternative is to summarize both raw number and population percentage results in terms of clutter (non-target) and target in a simple  $2 \times 2$  composite matrix (Table 2.3).

The strength of the confusion matrix is that it not only indicates how well the CS is doing over the entire data set, but it also gives clues as to where the errors are being made. Investigating

Table 2.3 Example Composite Confusion Matrix.  
Classified As (Reported)

Actual (Truth)		Clutter	Target
	Clutter	82 (91.1%)	8 (8.9%)
	Target	13 (14.4%)	77 (85.6%)

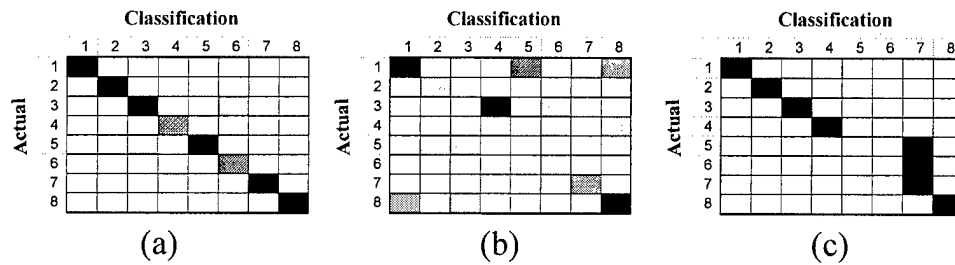


Figure 2.1 Typical gray level confusion matrices. (a) shows a good CS with few classification errors, (b) shows a poor CS with many classification errors, and (c) shows a CS with structured errors.

where these errors occur can be a useful method for determining which type of target data to collect if more data is considered necessary to better distinguish the target distribution from the clutter distribution and hence improve the CS performance. The drawbacks of the confusion matrix are that it is only a visualization of the raw data for one specific decision threshold and it does not provide a measure of effectiveness which could be used to compare various CSs.

The standard confusion matrix is not, necessarily, the best visualization tool available. Better means for visualizing the raw data to quickly identify the distribution of errors that a CS makes are available. Swingler [65] shows how it is possible to plot the confusion matrix using gray levels to indicate frequency as depicted in Figure 2.1. The darker the shading of a square in the grid, the more frequently the classifier produced an answer listed on the same column as the square when the correct answer was that denoted by the square's row. A near perfect classifier produces a confusion matrix with a very dark right hand diagonal and very pale entries elsewhere. These gray level confusion matrices enable the evaluator to quickly identify the distribution of errors that a classifier makes and thereby visualize its accuracy and simultaneously determine clues as to which aspect of the classifier's task needs improvement. For example, Figure 2.1 depicts three typical gray level confusion matrices. Gray level confusion matrices for competing classifiers could be compared side by side or a gray level matrix for the confusion matrix formed by computing the

difference between the competing classifiers' confusion matrices could be examined to determine visually where the two classifiers differ.

*2.2.2 Error Histograms.* For classifiers with several outputs or in situations where the size of the errors is more important than their type, an error histogram provides another quick method for visualizing the distribution of errors [65]. An error histogram shows the count of the frequency with which a classification error falls within a set of bandwidths, i.e., within a certain range of error sizes. These bandwidths or error sizes are the ranges of possible differences between the actual target class and the predicted class for each exemplar. For a classification probability score from zero to one, these bands must be split into a set of small bins. This error binning technique contrasts the setting of class thresholds used to classify the exemplars and generate the confusion matrix. For a simple two class confusion matrix, if the predicted classification score for a particular exemplar exceeds some preset threshold (e.g., 0.5), then that exemplar is classified as class 2. For the error histogram, the difference between a given exemplar's predicted classification probability and each actual target output class probability (e.g., if actual class is 2, then target probabilities are: 0.0 for Class 1 and 1.0 for Class 2) is used. A healthy classifier will show a peak at zero, quickly falling off as the number of errors of greater magnitude diminishes. For a data set with normally distributed noise, the error histogram should have the appearance of a normal distribution. Figure 2.2, shows an example of an error histogram for a notional classification problem. This histogram is constructed by splitting the real-valued classification errors (-1 to 1) into 21 bins and counting the number of errors in each bin. Since the majority of the errors are made in the small error bins, the notional error histogram signifies a healthy classifier.

*2.2.3 Error-Reject Curves.* Another technique used in pattern recognition is to allow classifiers to make *doubt* reports. Rather than making a firm classification, for example, of target or clutter, the classifier is permitted to identify exemplars which are too hard to classify, i.e., the classification output falls in a *gray* or uncertain area. These difficult exemplars are then rejected by

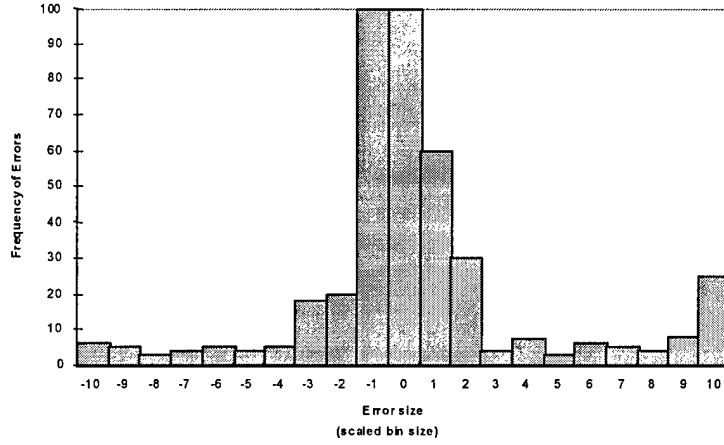


Figure 2.2 Error Histogram example for a well trained classifier

the classifier until further measurements can be made which permit a more definite classification or perhaps these difficult exemplars are passed on to a second classifier specifically designed to deal with the *gray* area of classification [60]. Using the doubt option, a loss function,  $L(k, l)$  can be defined as the loss incurred by making decision  $l$  if the true class is  $k$  (out of total of  $K$  classes). If every misclassification is equally serious, then the loss function is given by

$$L(k, l) = \left\{ \begin{array}{ll} 0 & \text{if } l = k \text{ (correct classification)} \\ d & \text{if } l = \mathcal{D} \text{ (classification in doubt)} \\ 1 & \text{if } l \neq k \text{ and } l \in \{1, \dots, K\} \text{ (incorrect classification)} \end{array} \right\} \quad (2.1)$$

where  $k = 1, \dots, K$  and  $l \in \{1, \dots, K\}$  is a reasonable choice [60]. The total risk for the optimal decision rule is called the Bayes risk ( $R$ ) and is defined by

$$R = p_{mc} + d \cdot p_d \quad (2.2)$$

where  $p_{mc}$  is the probability of misclassification or error,  $p_d$  is the probability of doubt, and  $d$  is the rejection threshold or the cost of being in doubt. The plot of  $p_{mc}$  versus  $p_d$  for varying  $d$ , is called the error-reject curve and is illustrated in Figure 2.3. The error-reject curve is a useful

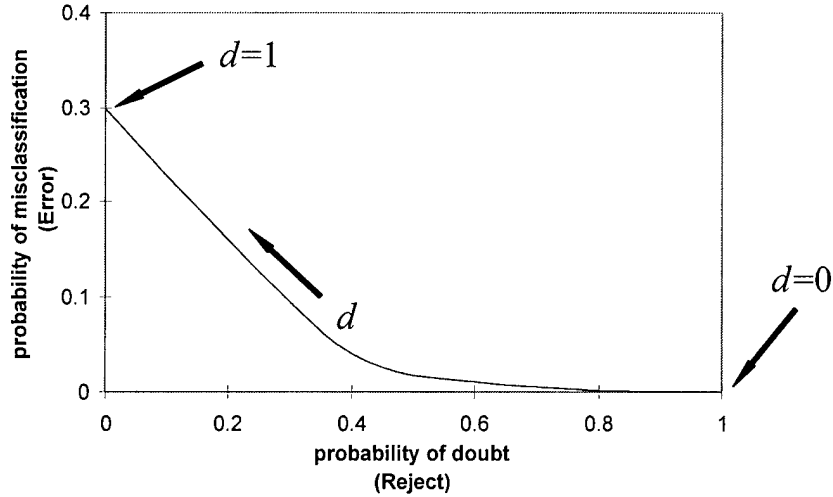


Figure 2.3 Error-Reject tradeoff curve.

performance tool for understanding the error-reject tradeoffs for a given classifier for the assumed cost  $d$  of rejecting data that is difficult to categorize [60]. Since the slope of the error-reject tradeoff curve is the value of the rejection threshold [60], the tradeoff is most effective for low levels of rejection and becomes less effective for high levels of rejection when the error rate is very low [22].

**2.2.4 Confidence Intervals.** A CS's performance is typically assessed using a set of probabilities. The most common performance measure used is classification accuracy ( $CA$ ) or the probability of success ( $p_S$ ), i.e., the probability of identifying targets and non-targets alike. For an ATR application [10,21,61], the typical probabilities used are the probability of detection ( $p_D$ ), probability of correct classification ( $p_{CC}$ ), probability of correct identification ( $p_{ID}$ ), and probability of false alarm ( $p_{FA}$ ). AFRL defines *correct detection* as correctly declaring that a target in a region of interest (ROI) is, in fact, a target. Conversely, a *false alarm*, or incorrect detection, occurs when the ATR declares clutter, such as trees, as a target. AFRL defines *correct classification*, as correctly classifying a detected target as a member of its actual target class regardless of the specific target type. For example, MSTAR is being designed to operate under realistic military scenarios, called extended operating conditions (EOC's), which include up to 20 specific target types in five different



Table 2.4 MSTAR EOC Class and Target Types

Class	Main Battle Tank (MBT)	Armored Personnel Carrier (APC)	Self-Propelled Gun (SPG)	Truck (T)	Mobile Missile Launcher (MML)
Target Type	T72 M1	BMP2 M2 M113 BTR60 BTR70	M109 M110	M548 M35 HMMWV	SCUD

classes. Thirteen (the types used in the first phase of the MSTAR program) of these 20 target types are shown in Table 2.4 [21].

If an M2 armored personnel carrier (APC) image is inputted and MSTAR reports APC as the image class, a correct classification is obtained, even if MSTAR incorrectly identifies the image type as a M113. *Correct Identification*, a subset of classification, is naming the specific alphanumeric target designator. For example, if a T72 main battle tank (MBT) image is inputted and MSTAR identifies the image as a T72, a correct identification [21] is obtained. The typical performance probabilities of interest ( $p_S, p_D, p_{CC}, p_{ID}, p_{FA}$ ) can be estimated as functions of the elements of the confusion matrix and the ground truth data [10,21]. The estimation equations are listed below along with sample calculations for the data in Table 2.1, where for illustration target 1 is assumed to be a T72, target 2 a M1, and target 3 a M2.

$$CA = \hat{p}_S = \frac{\text{number of target and clutter images correctly classified}}{\text{total number of target and clutter images}} = \frac{159}{180} = 88.3\% \quad (2.3)$$

$$\hat{p}_D = \frac{\text{number of target images declared as targets}}{\text{number of target images}} = \frac{77}{90} = 85.6\% \quad (2.4)$$

$$\hat{p}_{CC} = \frac{\text{number of correctly classified target images}}{\text{number of detected target images}}; \hat{p}_{CC}(\text{MBT}) = \frac{50}{77} = 64.9\% \quad (2.5)$$

$$\hat{p}_{ID} = \frac{\text{number of correctly identified target images}}{\text{number detected target images}}; \hat{p}_{ID}(T72) = \frac{24}{77} = 31.2\% \quad (2.6)$$

$$\hat{p}_{FA} = \frac{\text{number of clutter images declared as targets}}{\text{number of clutter images}} = \frac{8}{90} = 8.9\% \quad (2.7)$$

Often, performance measures such as the probabilities of success are reported as single numbers as calculated above. For example, an ATR designer might say that his system has a classification accuracy of 88.3 percent based on the probability of success estimated above. However, this is just a point estimate. Since the ATR is tested on a finite data set, the true classification accuracy is probably not 88.3 percent. Instead, the accuracy is more likely in some interval centered about the point estimate. For this example, a 95 percent confidence interval, assuming a binomial distribution for the number of successful classifications, is given by the interval [0.83 0.93]. In other words, if the ATR designer computed interval estimates from many different samples, then in the long run, he would expect about 95 percent of the intervals to include the true value for the accuracy of the ATR system. Hence, the confidence interval describes the experimental uncertainty in estimating the true ATR classification accuracy. ATRWG paper no. 88-006 [11] provides an excellent review of confidence intervals in ATR performance evaluation. The general procedure for construction of confidence intervals is to first postulate an underlying distribution. In ATR as in many CSs, the distributions which are probably of most interest are the Binomial, Poisson, and Gaussian. As an illustration, the concepts and calculation for a confidence interval for classification accuracy ( $CA$ ), i.e., the probability of success parameter,  $\hat{p}_S$ , is summarized below using the Binomial distribution to model the number of successful classifications.

Suppose an ATR is designed and then tested on an independent sample. For each image tested, there are two possible outcomes

$$\eta = \begin{cases} 0, & \text{if image incorrectly classified} \\ 1, & \text{if image correctly classified} \end{cases} \quad (2.8)$$

with associated probabilities:  $P(0) = 1 - p$  and  $P(1) = p$ , which means  $\eta$  is a Bernoulli random variable. For a series of these independent, identical trials, the Binomial random variable  $Y$  is the number of successful classifications in  $n$  trials, i.e.,  $Y = \text{binomial}(n, p)$ , where

$$p(y) = \binom{n}{y} p^y (1 - p)^{n-y} \quad (2.9)$$

and the expectation and variance of  $Y$  are given by

$$\mathbf{E}(Y) = np \text{ and } \mathbf{Var}(Y) = np(1 - p). \quad (2.10)$$

An unbiased estimate for  $p$ , the true classification success rate (probability of success), can be made using the definition in Equation 2.3 above as shown in Equation 2.11 below

$$\hat{p} = \frac{Y}{n}. \quad (2.11)$$

Now,  $\hat{p}$  is an unbiased estimator for  $p$  so

$$\mathbf{E}(\hat{p}) = \mathbf{E}\left(\frac{Y}{n}\right) = \frac{1}{n} \mathbf{E}(Y) = p. \quad (2.12)$$

Using the variance expression for  $\hat{p}$  yields

$$\mathbf{Var}(\hat{p}) = \mathbf{Var}\left(\frac{Y}{n}\right) = \frac{1}{n^2} \mathbf{Var}(Y) = \frac{p(1-p)}{n} \quad (2.13)$$

and the usual method of substituting sample values for unknown parameters in the expression for the variance, one can approximate  $(1 - \alpha)$  confidence intervals for  $\hat{p}$  as

$$\hat{p} \pm Z_{(1-\frac{\alpha}{2})} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad (2.14)$$

where the normal approximation is used (assuming large test sample size;  $n > 30$ ) for the binomial.

For the example in Table 2.3, with  $\alpha = .05$ , the following result referred to above is obtained

$$\hat{p} = 0.883 \pm 0.047 \quad \text{or} \quad 0.83 \leq \hat{p} \leq 0.93. \quad (2.15)$$

The strength in using a confidence interval is that it provides a quantifiable measure of the accuracy of the evaluation process. The confidence interval accounts for the sampling error of the testing experiment. When comparing the performance of various CSs tested under the same conditions, confidence intervals provide a simple measure of the variations in the performance results for the individual CSs.

One limitation of confidence intervals is that they can only indicate what can be expected in the future when one performs exactly the same test under the exact same conditions. For example, in their work [61], Ross *et al.* distinguish between data sets and conditions for an ATR application. They define a condition as a subset of a multi-dimensional space where the dimensions of the space are the specific conditions that may affect the performance of an ATR system. These dimensions can be grouped into those related to the target, environment, and sensor, as illustrated for synthetic aperture radar (SAR) ATR in Table 2.5.

Table 2.5 Example of SAR ATR dimensions by group

Group	Target	Environment	Sensor
Dimensions	# targets # variants # configurations	obscuration background	radar frequency depression angle

In general, it may not be possible to define this multi-dimensional space, but for a particular ATR test or application, it is often possible to explicitly list the dimensions and their range. Ross *et al.* [61] define the conditions for which the dimensions take on all possible values in their range as the universal set. There are four subsets of special interest within this universal set:

1. *operational conditions*-ranges of conditions ATR expected to handle operationally;
2. *testing conditions*-ranges of conditions used in testing ATR;
3. *training conditions*-ranges of conditions used for training ATR;
4. *modeled conditions*-ranges of conditions modeled on-line for model-driven ATR.

*Operational conditions* are the conditions under which the ATR user expects the ATR to function properly. Ross *et al.* [61] warn that the user may assume that the reported ATR performance is over these operational conditions. However, this may often not be correct. The *testing conditions* are typically a subset of the operational conditions, but depending upon the data available, might expand on the operational conditions in some dimensions. *Training conditions* are more than just the exact conditions of the various training data. Ross *et al.* [61] define the training conditions to include all of those conditions that produce images that are *similar* to training images. Images are considered *similar* if a simple non-feature-based comparison between each image and the target image result in mean-square-errors small enough to allow accurate recognition of both images. The training conditions are typically a proper subset of the testing conditions. For a model-driven ATR, the *modeled conditions* include the conditions modeled on-line. For example, the on-line model may be capable of handling 10 target types and 5 to 15 GHz radar frequency, but cannot deal with obscuration. The modeled conditions are then 10 target types and 5 to 15 GHz radar

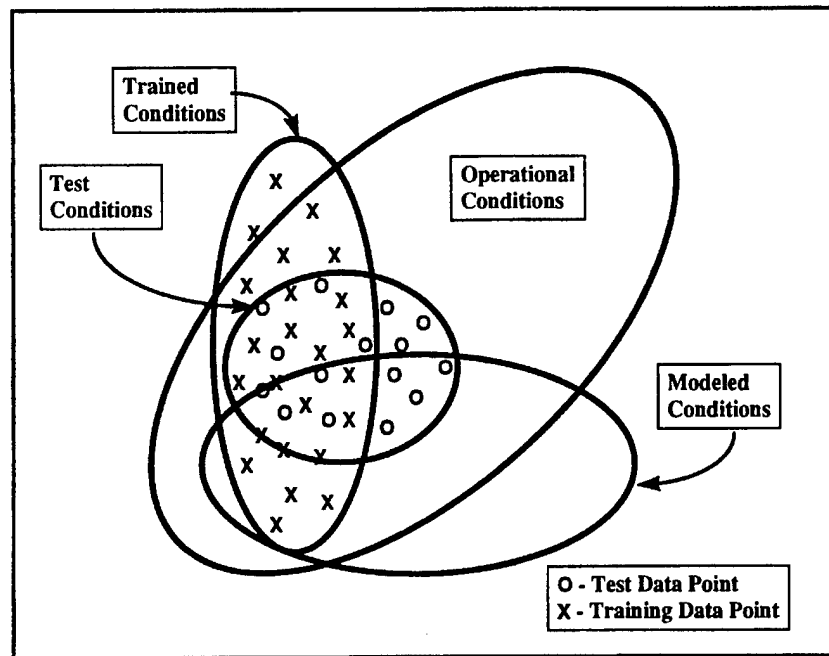


Figure 2.4 Venn diagram showing the relationship between modeled, training, and testing conditions [61].

frequency. Data sets are finite collection of individual data points in the space of conditions, as illustrated in Figure 2.4.

Since ATR systems are operated and tested under these conditions, four performance measures can be defined using these conditions [61]:

1. *Accuracy*-how well an ATR performs under its training conditions;
2. *Robustness*-how well an ATR performs outside its training conditions;
3. *Extensibility*-how well a model-based ATR performs outside training conditions, but within its modeled conditions;
4. *Utility*-how well an ATR performs under operational conditions.

The relationship between these four performance measures and the conditions used to define and test them are illustrated in Figure 2.5.

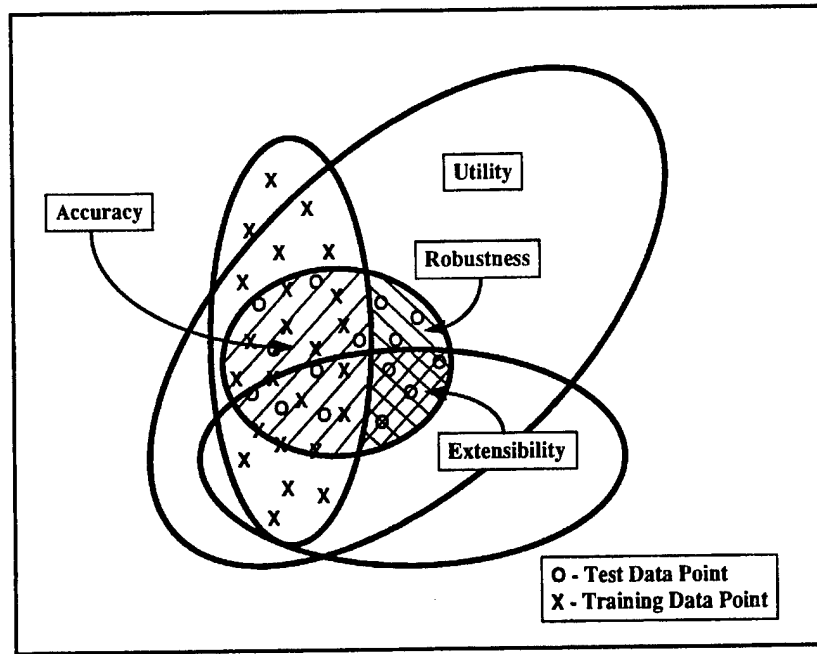


Figure 2.5 Venn diagram showing the relationship between accuracy, robustness, extensibility, and utility [61].

In order for performance results to be useful indicators of future performance, the future conditions must be the same as the testing conditions. For example, Ross et al. [61] warn that unless the training conditions are the same, or nearly the same, as the operational conditions of the fielded ATR system, then the accuracy of the system will not be a good indicator of system performance outside the training conditions.

An assumption made in computing confidence intervals for ATR performance is that the test sample is a random sample, i.e., the *a priori* probabilities of occurrence  $\pi_i$  for each class  $i$  of the test sample are unknown [11]. For this case of *random sampling*, test points are considered to be randomly generated by a “pattern source” according to the *a priori* probabilities of occurrence  $\pi_i$  [39]. An ATR evaluator takes a test point or sample pattern from this “pattern source,” identifies it, and then lets the ATR in question attempt identification. This experiment is repeated  $n$  times, resulting in  $Y$  samples or test points which have been successfully classified. Estimates for

the mean and variance for classification accuracy for the case of *random sampling* are then given by Equation 2.11 and 2.13 above.

For the alternative case of *selective sampling*, the assumption is made that the *a priori* probabilities of occurrence  $\pi_i$  for each class  $i$  (of  $k$  total classes) of the test sample are known [39]. To take advantage of this knowledge, the ATR evaluator takes  $n_i$  samples from each class  $i$  such that

$$\frac{n_i}{n} = \pi_i \quad (2.16)$$

where  $n$  is the total number of samples. The ATR in question is then again allowed to attempt recognition of these test samples, resulting in  $Y_i$  samples from class  $i$  being correctly classified. Then for this case of *selective sampling*, the estimates for the mean  $\hat{p}'$  and variance  $\text{Var}(\hat{p}')$  for classification accuracy are given by

$$\hat{p}' = \frac{\sum_{i=1}^k Y_i}{n} \quad (2.17)$$

$$\text{Var}(\hat{p}') = \frac{1}{n} \sum_{i=1}^k \pi_i \frac{Y_i}{n_i} \left( 1 - \frac{Y_i}{n_i} \right). \quad (2.18)$$

Highleyman [39] shows that the variance in the case of *selective sampling* is smaller than the variance in the case of *random sampling*. The result is that the confidence intervals for classification accuracy in the case of *selective sampling* are tighter than the *random sampling* confidence intervals usually computed.

A drawback to using independent samples, whether it be *random* or *selective sampling*, to test a CS such as an ATR system is that the training data is not typically exploited to the fullest extent possible. Training data is used to train the ATR and then it is discarded. As long as the training data is representative of the operational conditions deemed important (i.e., training conditions the



same or nearly the same as the operational conditions), then the training data can itself be used to obtain estimates and confidence intervals for the classification error using techniques that correct for the inherent bias in using the training data for testing. James [45] provides an excellent review of two of these methods: the leaving-one-out and jackknife methods.

One must also be aware that confidence intervals should not be used out of context to predict the range of future values for performance measures when the CS is operated over the full range of its dimensions. For example, if an ATR designer is interested in knowing the range of detection rates that can be expected from his system when it operates over a range of dimensions, such as various aspect angles, this information can only be obtained by performing a series of tests over the desired range of interest. Finally, as with confusion matrices, confidence intervals are based on one specific decision threshold. If the decision threshold is changed, a new result for the performance parameter may result.

*2.2.5 Hypothesis Testing.* The procedure of hypothesis testing is directly related to the use of confidence intervals. An operational CS must satisfy certain performance specifications. For example, AFRL requires MSTAR to realize a probability of detection  $p_D \geq 0.9$  with 95% confidence [21]. Hypothesis testing allows the program evaluator to infer the detection performance for a certain configuration of MSTAR on the entire population of target images after testing on a limited sample [48]. The null and alternative hypotheses are:

$$H_0 : p_D \geq 0.9 \text{ (ATR meets specification)} \quad (2.19)$$

$$H_1 : p_D < 0.9 \text{ (ATR fails to meet specification)}$$

Using the binomial model in a similar fashion as in Section 2.2.4 for confidence intervals, Equations 2.12 and 2.13 above are applied (for large  $n$ ) to derive the test statistic for this hypothesis test

$$Z_0 = \frac{\hat{p}_D - p_0}{\sqrt{\frac{\hat{p}_D(1-\hat{p}_D)}{n}}} \approx N(0, 1) \quad (2.20)$$

where  $p_0 = 0.90$  is the required AFRL specified value for detection performance. Consider the simple notional example from Table 2.3 again, which consists of a test sample of  $n = 90$  target images where  $Y = 77$  images are correctly identified as targets. The point estimate of  $\hat{p}_D = \frac{77}{90} = 0.856$  would incorrectly suggest that the ATR does not satisfy AFRL's requirement [48]. Applying the test statistic in Equation 2.20 to the one-sided hypothesis test (Equation 2.19) above (with  $\alpha = 0.05$ ) yields

$$Z_0 = -1.189 > -Z_{1-\alpha} = -1.645 \quad (2.21)$$

which implies there is no statistical evidence for rejecting the null hypothesis that the ATR meets the required specification.

The strength of hypothesis testing is that it provides a formal approach to the statistical evaluation of CSs [48]. However, hypothesis testing shares the same weaknesses and limitations of confidence intervals mentioned above. In addition, the distribution postulated for the interested performance parameter carries with it certain assumptions. For example, the confidence interval and hypothesis testing methods based on the Binomial distribution assumes a constant probability of success (or detection) and a constant variance for all observations. Since AFRL wants the data for MSTAR to represent the full range of aspect and depression angles for all of the targets, performance parameters, such as probability of correct identification  $p_{ID}$ , may differ for different targets at different angles [21]. Catlin et al. [21] provide a derivation of a confidence interval which accounts for variation in the scenario  $p_{ID}$ 's.

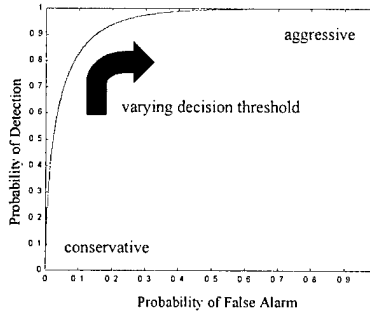


Figure 2.6 Typical ROC curve.

**2.2.6 ROC Curves.** Receiver operating characteristic (ROC) curves are commonly used for summarizing the performance of imperfect diagnostic systems, especially in automatic target recognition (ATR) and in biomedical research, when classification accuracy alone is not sufficient. A ROC curve is the graph of a relation which summarizes the possible performances of a signal detection system faced with the task of detecting a signal (target) in the presence of background noise (clutter). This relation is usually used to relate the detection or “hit” rate (probability of detection, i.e., probability of true positive) to the false alarm rate (probability of false alarm, i.e., probability of false positive) as an internal decision threshold is varied. For a typical ROC curve, shown in Figure 2.6, the decision threshold is varied from a very conservative value, i.e., a value that results in zero detection rate and zero false alarm rate, to a very aggressive value, i.e., a value that results in 100% detection rate and 100% false alarm rate.

In order to describe the basic principles behind standard ROC curves, consider the simple two-class problem ( $C_1$  is the clutter or non-target class and  $C_2$  is the target class) with a single variable or feature  $z \in \mathbb{R}$  depicted in Figure 2.7. Let  $Z$  be a real-valued random variable and let  $p(z)$  be its probability density function (pdf). Then the conditional pdf for each class will be  $p(z|C_1)$  and  $p(z|C_2)$ . That is, the clutter pdf,  $p(z|C_1)$ , is the conditional pdf representing the distribution of clutter objects (class 1) while the target pdf,  $p(z|C_2)$ , is the conditional pdf representing the distribution of target objects (class 2). Since the choice of scale for the  $z$ -axis is arbitrary and

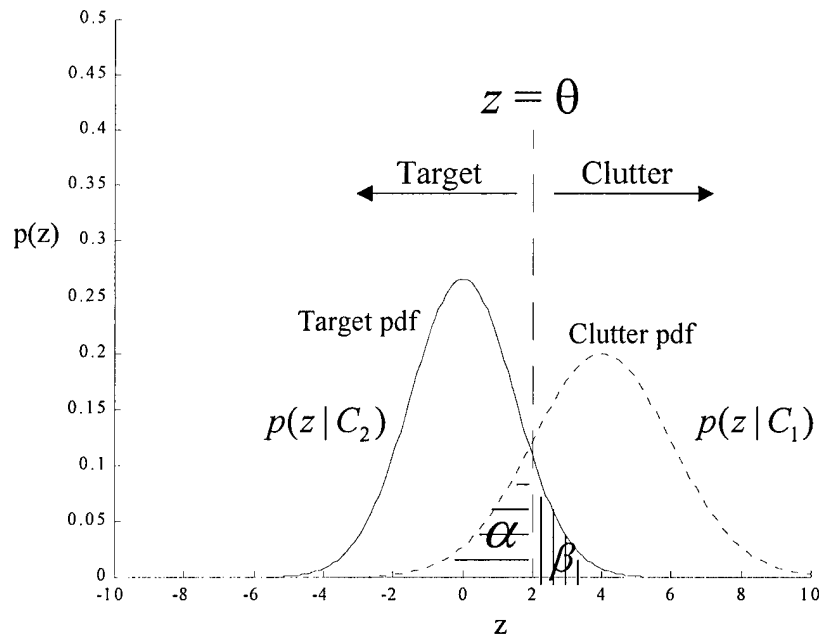


Figure 2.7 Target and clutter pdfs for a two class problem. Target:  $N(0, 1.5)$ ; Clutter:  $N(4, 2)$ .

is easily transformed, let lower values of  $z$  equate to stronger indications of target, while higher values of  $z$  equate to stronger indications of clutter. A given decision threshold boundary,  $z = \theta$ , then, partitions the feature axis into two regions, target  $(-\infty, \theta)$  and clutter  $(\theta, \infty)$ , resulting in two types of errors shown as horizontal ( $\alpha$ ) and vertical ( $\beta$ ) cross-hatched areas under the two distribution curves in Figure 2.7:

1. Type I Error ( $\alpha$ ): Misclassifying an actual clutter object as a target (False Positive, FP, or False Alarm, FA)
2. Type II Error ( $\beta$ ): Misclassifying an actual target object as clutter (False Negative, FN).

For a single feature problem, the clutter and target distributions can be more complex than the simple distributions in Figure 2.7. For many features, the class distributions will be even of greater complexity. However, a transformation can always be made to a simple one dimensional space  $Z$  where  $Z$  is the real-valued random variable representing the strength of conviction for the

non-target or clutter. Therefore, the conditional probabilities,  $P_{FP}$  and  $P_{FN}$ , corresponding to the two types of errors described above can be defined as

$$\alpha = P_{FP}(\theta) = \Pr(\{z < \theta \mid C_1\}) \quad (2.22)$$

$$\beta = P_{FN}(\theta) = \Pr(\{z > \theta \mid C_2\}). \quad (2.23)$$

Associated with these two probability errors are their complementary probabilities of correct classification,

$$P_{TN}(\theta) = \Pr(\{z > \theta \mid C_1\}) \quad (2.24)$$

$$P_{TP}(\theta) = \Pr(\{z < \theta \mid C_2\}). \quad (2.25)$$

The interrelationships among these probabilities and the various terminologies used in ATR, statistics, and medicine to describe them are shown in Figure 2.8. Because of the interrelationships among the probabilities, it is only necessary to indicate a single (specificity, sensitivity) pair to describe the performance of a pattern recognition algorithm for a particular decision threshold. These (specificity, sensitivity) probability pairs vary as the decision boundary shifts along the feature axis. For example, if the decision boundary  $\theta$  is shifted to the right in Figure 2.7,  $\beta$  decreases while  $\alpha$  increases. A relationship exists between these two error probabilities. In ROC analysis, the decision threshold  $\theta$  is purposefully varied over all possible  $\theta$  values to show this relationship in the form of a ROC curve. Figure 2.9, illustrates the generation of a ROC curve for the case of one-dimensional normal distributions for both the clutter and target distributions.

		Classified As (Reported):	
		Clutter- C (Normal-N) $H_0$	Target-T (Abnormal-A) $H_1$
Actual (Truth)	Clutter- c (Normal- n) $H_0$	True Neg Specificity Confidence (1- $\alpha$ ) $P_{TN}$ $P(C c)$ $P(N n)$	False Pos 1-Specificity Level of Sig $\alpha$ $P_{FP}$ ( $P_{FA}$ ) $P(T c)$ $P(A n)$
	Target- t (Abnormal- a) $H_1$	False Neg 1-Sensitivity 1-Power $\beta$ $P_{FN}$ $P(C t)$ $P(N a)$	True Pos Sensitivity Power (1- $\beta$ ) $P_{TP}$ ( $P_D$ ) $P(T t)$ $P(A a)$

Figure 2.8 Relationships and terms associated with ROC analysis (adapted from Hildebolt *et al.*, 1991 [40]).

ROC curves have their foundation in statistical decision theory [68] and were originally developed as tools for electronic signal detection [58]. ROC analysis has been extensively applied to human perception and decision-making problems [32] and is also commonly used in biomedical research [51]. For an in-depth technical discussion of ROC curves, consult Egan [27] and Swets and Pickett [64]. Han and Clark [37] provide a good introduction of ROC analysis applied to the ATR problem for a simple, single decision threshold parameter problem as described above. Irving and Wissinger [43] describe how to generate ROC curves for MSTAR when multiple decision threshold parameters are required. Of special significance is their approach for choosing a single set of threshold parameter settings for processing region of interests (ROIs) so that the ROIs need only be pushed through the ATR system once and that the average run-time per chip is no more than 30 minutes. The results from this run of a single set of threshold settings are rich enough that ROC curves can be generated by post processing the data. Since many ATR designs require the X-axis of the ROC curves to be in units of false alarms per kilometer square, Jachimczyk's work [44] is a good illustration of how this conversion is made.

The major strength of ROC curves in CS evaluation is that rather than reporting the system's performance in terms of a simple target detection *batting average* [51] for a specific decision

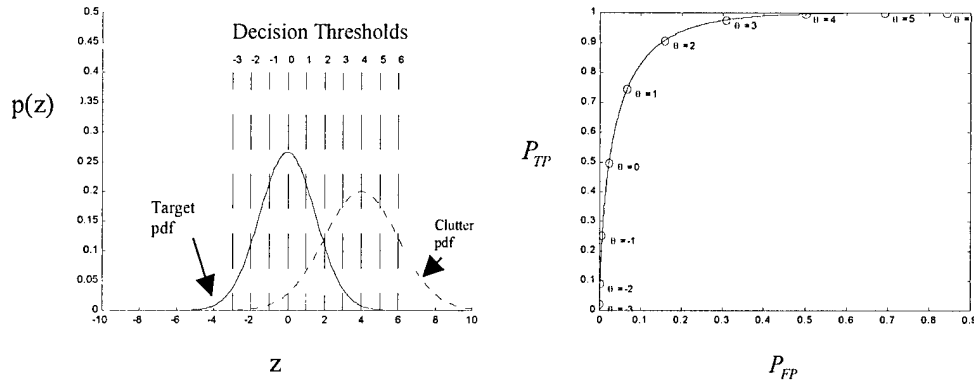


Figure 2.9 Generation of ROC curve (right) using two 1-D normal distributions (left). Target:  $N(0, 1.5)$ ; Clutter:  $N(4, 2)$ .

threshold, ROC curves enable performance reporting in terms of a pair of related indices (detection probability, false alarm probability) for varying decision thresholds. ROC curves provide a means for characterizing and qualitatively comparing CS designs. One CS design can visually be seen as *better* than an alternate design if its associated ROC curve is higher than its competitor's ROC curve as shown in Figure 2.10.

### 2.3 Performance Comparison

**2.3.1 Comparison of Confusion Matrices.** One technique [45] used in pattern recognition is to compare the form of the resulting confusion matrix from a classification problem to the confusion matrix forms for two simple classification rules where the prior probabilities are known. For example in the two class case where  $\pi_1$  and  $\pi_2$  are the a priori probabilities for the two classes, the confusion matrix takes on two distinctive forms based on whether a random or a priori assignment classification rule is used. If each exemplar is randomly assigned as it occurs to any of the two classes (random assignment), then the population confusion matrix is given by Table 2.6 and the total classification error is 0.5. If the exemplars are assigned randomly as before, but with the

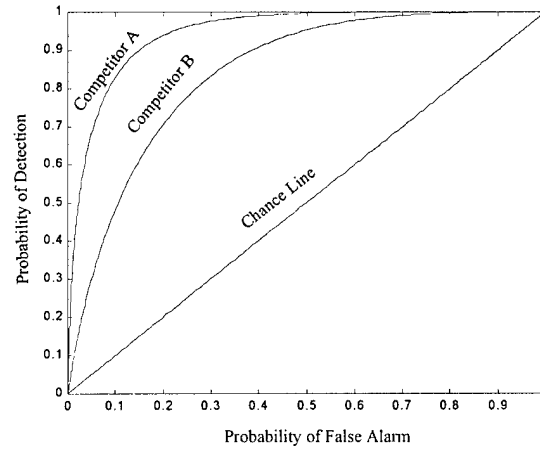


Figure 2.10 Qualitative comparison of ROC curves. Competitor A is *better* than Competitor B.

Table 2.6 Random assignment confusion matrix.  
Classified As (Reported)

Actual (Truth)		Class 1	Class 2	Prior Prob
	Class 1	$\frac{\pi_1}{2}$	$\frac{\pi_1}{2}$	$\pi_1$
	Class 2	$\frac{\pi_2}{2}$	$\frac{\pi_2}{2}$	$\pi_2$
	Assign Prob	$\frac{1}{2}$	$\frac{1}{2}$	

a priori probabilities  $\pi_1$  to class 1 and  $\pi_2$  to class 2 (a priori assignment), then the population confusion matrix is given by Table 2.7 and the total classification error is  $2\pi_1\pi_2$ .

These two assignment rule cases indicate what sort of errors and pattern of errors can be achieved *without really trying* [45]. The random assignment case corresponds to using no information about the class populations while the a priori assignment case uses only the a priori probabilities. These cases with their corresponding population confusion matrices shown in Table 2.6 and Table 2.7 and associated errors can be used as a baseline against which the performance of any classifier is judged. A Chi square ( $\chi^2$ ) statistic can be used as a measure of how far the observed

Table 2.7 A priori assignment confusion matrix.  
Classified As (Reported)

Actual (Truth)		Class 1	Class 2	Prior Prob
	Class 1	$\pi_1^2$	$\pi_1 \cdot \pi_2$	$\pi_1$
	Class 2	$\pi_1 \cdot \pi_2$	$\pi_2^2$	$\pi_2$
	Assign Prob	$\pi_1$	$\pi_2$	



confusion matrix differs from the expected confusion matrices for either the random or a priori assignment cases. The statistic  $\chi^2$  is computed in the usual way

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \quad (2.26)$$

where  $O_i$  and  $E_i$  are the respective observed and expected frequencies for the  $i^{th}$  cell of the raw numerical confusion matrix with a total of  $k$  cells ( $k = G^2$ , where  $G$  = number of classes) and the error rates of either of the two assumed random assignment cases are used to compute the  $E_i$ . If the  $O_i$  are based on an independent sample, then the statistic does indeed follow a distribution with  $(G - 1)^2$  degrees of freedom as the sample size increases [45]. A comparison of  $\chi^2$  for two competing classifiers can determine which classifier has a pattern of errors further removed from the assumed random assignment case and therefore, which classifier is using more information than random chance. In order to directly compare two competing classifiers, a  $\chi^2$  statistic could be computed using the numbers from one of the confusion matrices as the expected frequencies and the numbers from the other classifier's confusion matrix as the observed frequencies. The disadvantage of this comparison technique in practice, is that the  $\chi^2$  statistic becomes a function of the choice of the nominal classifier for the expected frequencies as well as the particular decision threshold for which the confusion matrix is computed.

Instead of comparing the entire confusion matrix of one classifier to another, a row by row comparison can be made between competing classifiers [57]. For each row of a confusion matrix each element of that row is conditioned the same. For a given row  $i$  for example, each element  $f_{ij}$ , representing the number of exemplars classified as class  $j$  given its actual class is  $i$ , is drawn from the conditional probability  $\Pr(\text{classified } j \mid \text{actual class } i)$ . To compare a row from a confusion

matrix between  $m$  competing classifiers for a  $k$  class problem, the  $\chi^2$  test statistic is given by

$$\chi^2 = n \left( \sum_{i=1}^m \sum_{j=1}^k \frac{f_{ij}^2}{n_i C_j} - 1 \right) \quad (2.27)$$

where  $n_i$  = row  $i$  totals,  $n = \sum_{i=1}^m n_i$ , and  $C_j = \sum_{i=1}^m f_{ij}$ . This  $\chi^2$  statistic should follow a  $\chi^2$  distribution with  $(k-1)(m-1)$  degrees of freedom.

### 2.3.2 Comparison of Classifiers Using Hypothesis Testing.

**2.3.2.1 Non-sequential Hypothesis Testing.** In order to compare two CSs, one can either decide in advance the number of images (or trials) for testing each system or one can have the testing procedure decide *on the fly* if more trials are needed to differentiate between the performance of the systems. In Section 2.3.2.1 the former procedure, called non-sequential testing, is considered, while in Section 2.3.2.2, the latter, called sequential testing, is discussed.

When comparing two CSs using non-sequential testing [21], one can either compare the confidence intervals for some performance measure  $p$  calculated via Equation 2.14 for both systems, or the confidence interval for the performance difference can be computed. Assuming equal sample sizes  $n$  ( $n = n_1 = n_2$ ), this difference interval can be calculated [29] using classical inferential statistics for large  $n$  ( $n > 30$ ) as:

$$(\hat{p}_2 - \hat{p}_1) \pm Z_{1-\alpha} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1) + \hat{p}_2(1-\hat{p}_2)}{n}} \quad (2.28)$$

with the associated test statistic,

$$Z_0 = \frac{\hat{p}_2 - \hat{p}_1}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1) + \hat{p}_2(1-\hat{p}_2)}{n}}} \quad (2.29)$$

for the following null and alternative hypotheses:

$$\begin{aligned} H_0 : \quad p_1 &\geq p_2 \\ H_1 : \quad p_1 &< p_2 \end{aligned} \tag{2.30}$$

The advantage of using the difference interval (or its associated hypothesis test) over the comparison method of two separate intervals, is that a superior system can be chosen with fewer samples. However, difference intervals still need large sample sizes when the difference is small [21]. Another problem with the classical approach above is that the assumption must be made that the probability  $p$  does not vary from trial to trial. As mentioned in Section 2.2.5 above, some probability measures may violate this assumption, which means that the classical approach is not truly valid in all applications. Wald's work [68] provides an exact, but cleverly simple method for probability comparison using a non-sequential testing procedure which allows for the variation of probabilities from trial to trial.

To illustrate Wald's procedure [68], consider again an ATR system which is designed and then tested on an independent sample. For each target tested, there are two possible outcomes:

$$\eta = \begin{cases} 0 & \text{if target image incorrectly classified as clutter} \\ 1 & \text{if target image correctly classified as target} \end{cases} \tag{2.31}$$

The results for each target test image using Equation 2.31 for two different ATR systems are arranged in pairs in the order observed. Define  $t_1$  as the number of pairs (1, 0) where ATR number 1 was successful at detecting a target image, while ATR number 2 was unsuccessful. Similarly, define  $t_2$  as the number of pairs (0, 1) where ATR number 2 was successful, but ATR number 1 was unsuccessful. Considering only the ordered pairs (1, 0) and (0, 1), the hypothesis tests in Equation

2.30 are equivalent to

$$\begin{aligned} H_0 : \quad p &\geq \frac{1}{2} \\ H_1 : \quad p &< \frac{1}{2} \end{aligned} \tag{2.32}$$

where  $p$  is the probability that any ordered pair  $(a, b)$  is equal to  $(0, 1)$  and is given by:

$$p = \frac{(1 - p_1) p_2}{p_1(1 - p_2) + p_2(1 - p_1)} \tag{2.33}$$

The test statistic for the equivalent hypothesis tests in Equation 2.32 is simply the number  $t_2$  of observed ordered pairs  $(0, 1)$ . The null hypothesis, that  $p_1$  is better than  $p_2$ , is rejected only if,  $t_2 \geq T$ , where the value of  $T$ , for a given level of significance  $\alpha$ , is given by the binomial distribution with  $p = \frac{1}{2}$  :

$$\Pr(t_2 \geq T) = \sum_{i=0}^T \binom{t}{i} p^i (1-p)^{t-i} = 1 - \alpha \quad \text{where } t = t_1 + t_2 \tag{2.34}$$

**2.3.2.2 Sequential Hypothesis Testing.** The main problem with non-sequential testing, which is the common practice in most CS evaluation, is that more test samples are used on average than are really necessary if sequential testing was used instead. Catlin et al. [21] provide an excellent in-depth description and application of Wald sequential testing [68]. This section will provide a brief overview of the Wald sequential testing method and highlight the application results of Catlin et al.

Wald sequential testing is a logical extension of Wald's exact non-sequential method discussed in Section 2.3.2.1 above. Consider the same ATR detection example from the previous section. The Wald sequential test is then based on the *efficiencies* of the two competing ATRs, where Wald

defines efficiency  $k$  as

$$k = \frac{p}{(1-p)} \quad (2.35)$$

such that  $p$  is the true probability of success, which for this example is the true probability of detection. The relative superiority of ATR number 2 over ATR number 1 can be measured by the ratio ( $u$ ) of the efficiencies of two systems:

$$u = \frac{k_2}{k_1} = \frac{p_2(1-p_1)}{p_1(1-p_2)} \quad (2.36)$$

To implement the test, one must first set four parameters, which reflect the precision ( $u_0, u_1$ ) required and the risks ( $\alpha, \beta$ ) tolerated. Wald explains the procedure for choosing  $u_0$  and  $u_1$  in terms of manufacturing processes:

...select two values of  $u, u_0$ , and  $u_1$  say ( $u_0 < u_1$ ), such that the rejection of process 1 in favor of process 2 is considered an error of practical importance whenever the true value of  $u \leq u_0$ , and the maintenance of process 1 is considered an error of practical importance whenever  $u \geq u_1$ . If  $u$  lies between  $u_0$  and  $u_1$ , the manufacturer does not particularly care which decision is made. [68]

Since  $u$ -space (the set of  $u$  values for  $u_0$  and  $u_1$ ) does not clearly indicate precision, Catlin et al. recommend investigating  $u$  values for various  $p_1$  and  $p_2$  values [21]. Table 2.8 is an example of  $u$  values for various  $p_1$  and  $p_2$  values. Since the most interesting differences in  $p_1$  and  $p_2$  values occurs when both are near the desired performance value  $p_0$ , a table of  $u$  values for various  $p_1$  and  $p_2$  values in the neighborhood of  $p_0$  provides a good starting point for the required level of precision [21]. Table 2.9 is an example of  $u$  values in the neighborhood of  $p_0 = 0.70$ .

As an illustration of the selection of  $u_0$  and  $u_1$  using Table 2.9 above, consider the following example. Suppose ATR number 1 is currently being used with probability of detection  $p_1$ . If ATR number 1 is preferred when  $p_1 - p_2 > 0.03$ , then possible  $u_0$  values are 0.86 or 0.87. If ATR

Table 2.8 Values of  $u$  for the full range of possible  $p_1$  and  $p_2$  values [21].

$p_1/p_2$	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
<b>0.10</b>	0.00	1.00	2.25	3.86	6.00	9.00	13.5	21.0	36.0	81.0
<b>0.20</b>	0.00	0.44	1.00	1.17	2.67	4.00	6.00	9.33	16.0	36.0
<b>0.30</b>	0.00	0.26	0.58	1.00	1.56	2.33	3.50	5.44	9.33	21.0
<b>0.40</b>	0.00	0.17	0.38	0.64	1.00	1.50	2.25	3.50	6.00	13.5
<b>0.50</b>	0.00	0.11	0.25	0.43	0.67	1.00	1.50	2.33	4.00	9.00
<b>0.60</b>	0.00	0.07	0.17	0.29	0.44	0.67	1.00	1.56	2.67	6.00
<b>0.70</b>	0.00	0.05	0.11	0.18	0.29	0.43	0.64	1.00	1.71	3.86
<b>0.80</b>	0.00	0.03	0.06	0.11	0.17	0.25	0.38	0.58	1.00	2.25
<b>0.90</b>	0.00	0.01	0.03	0.05	0.07	0.11	0.17	0.26	0.44	1.00
<b>1.00</b>	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 2.9 Values of  $u$  for selected  $p_1$  and  $p_2$  values from 0.66 to 0.75 [21].

$p_1/p_2$	0.66	0.67	0.68	0.69	0.70	0.71	0.72	0.73	0.74	0.75
<b>0.66</b>	1.00	1.05	1.09	1.15	1.20	1.26	1.32	1.39	1.47	1.55
<b>0.67</b>	0.96	1.00	1.05	1.10	1.15	1.21	1.27	1.33	1.40	1.48
<b>0.68</b>	0.91	0.96	1.00	1.05	1.10	1.15	1.21	1.27	1.34	1.41
<b>0.69</b>	0.87	0.91	0.95	1.00	1.05	1.10	1.16	1.21	1.28	1.35
<b>0.70</b>	0.83	0.87	0.91	0.95	1.00	1.05	1.10	1.16	1.22	1.29
<b>0.71</b>	0.79	0.83	0.87	0.91	0.95	1.00	1.05	1.10	1.16	1.23
<b>0.72</b>	0.75	0.79	0.83	0.87	0.91	0.95	1.00	1.05	1.11	1.17
<b>0.73</b>	0.72	0.75	0.79	0.82	0.86	0.91	0.95	1.00	1.05	1.11
<b>0.74</b>	0.68	0.71	0.75	0.78	0.82	0.86	0.90	0.95	1.00	1.05
<b>0.75</b>	0.65	0.68	0.71	0.74	0.78	0.82	0.86	0.90	0.95	1.00

number 2 with probability of detection  $p_2$  is the preferred classifier when  $p_2 - p_1 > 0.03$ , then possible  $u_1$  values range from 1.15 to 1.17. This is an example of an unfair comparison, which is common in manufacturing. When competing ATRs are compared fairly, as is the current practice in the MSTAR case, the selection of  $u_0$  and  $u_1$  is simplified. For fair comparisons,  $u_0$  and  $u_1$  are reciprocals, i.e.,  $u_0 = \frac{1}{u_1}$ .

For statistical tests,  $\alpha$  and  $\beta$  are the desired significance level and 1 minus the power of the test respectively. The significance level  $\alpha$  is the probability of a Type I error and  $\beta$  is the probability of a Type II error (see page 2-18 above). Wald also defines  $\alpha$  and  $\beta$  as risk tolerances for his sequential test method:

The probability of rejecting process 1 should not exceed a pre-assigned value  $\alpha$  whenever  $u \leq u_0$ , and the probability of maintaining process 1 should not exceed a pre-assigned value  $\beta$  whenever  $u \geq u_1$ . [21]

Wald is using the parameters  $\alpha$  and  $\beta$  in dual roles in his sequential testing method. These parameters represent both risk tolerances for choosing the wrong process, as well as test significance and power [21].

As in Wald's non-sequential test, the null and alternate hypotheses are given by Equation 2.30 and the test statistic is again  $t_2$ , the number of observed ordered pairs  $(0, 1)$ , representing the number of trials where there was a successful target detection by ATR number 2 and failure of ATR number 1. But, this time for the sequential test, instead of comparing  $t_2$  to just one critical value,  $t_2$  is compared to two critical values: *the acceptance and rejection numbers*.

$$\begin{aligned} \text{acceptance number} \quad a_t &= \frac{\log \frac{\beta}{1-\alpha}}{\log u_1 - \log u_0} + (t_1 - t_2) \frac{\log \frac{1+u_1}{1+u_0}}{\log u_1 - \log u_0} \quad (\text{lower bound}) \\ \text{rejection number} \quad r_t &= \frac{\log \frac{1-\beta}{\alpha}}{\log u_1 - \log u_0} + (t_1 - t_2) \frac{\log \frac{1+u_1}{1+u_0}}{\log u_1 - \log u_0} \quad (\text{upper bound}) \end{aligned} \quad (2.37)$$

If  $t_2$  falls below the acceptance number for any value of  $t = t_1 + t_2$ , the null hypothesis that ATR number 1 is better than ATR number 2 is accepted. If  $t_2$  exceeds the value for the rejection number, the null hypothesis is rejected and the conclusion is that ATR number 2 is better than ATR number 1. If  $t_2$  remains between these bounds, testing is continued.

In their work, Catlin et al. [21] applied the Wald sequential test methodology to actual data to compare the probability of identification ( $p_{ID}$ ) performance measure for different configurations of the MSTAR System. They also embedded the Wald sequential test methodology inside a multiple sequential rejective Bonferroni procedure for the multiple pairwise comparison of more than two ATR systems. The Wald test required only an average of about one sixth as many samples as confidence intervals to choose the superior of two system configurations, and about one fifth as many samples as the non-parametric method of ranking and selection. In a four system comparison with simulated data, the embedded Wald test typically needed only one third as many samples as

multiple pairwise confidence intervals to detect specified differences between system  $p_{ID}$ 's, and about one half as many samples as required by ranking and selection.

The results of Catlin et al. emphasize the sample size savings advantage of the Wald sequential test methodology. Since image data collection can be very expensive, CS designers prefer to compare CS systems with the smallest number of sample images to choose one system as statistically significantly better than another. The Wald sequential test methodology appears to be a good method for these cost conscious designers.

The only limitations of the Wald methodology as presented by Catlin, are the use of a fair comparison procedure and the equal treatment or weighting of targets in her MSTAR application. If an unfair comparison is required, such that performance is weighted by cost or another quantitative measure, the methodology can be compensated by adjusting the likelihood function and applying Wald's sample size formulas for the general case. If the best mission specific ATR system is desired, target sampling can be modified to favor the more likely targets in the mission scenario. Competing systems can then be compared on their performance for targets which are representative of the specific mission.

*2.3.3 Comparison of ROC Curves.* ROC analysis is widely accepted and used in medical applications to evaluate the accuracy of diagnostic and prognostic technologies. However, instead of just reporting qualitative comparisons between ROC curves of competing systems, as discussed in Section 2.2.6, quantitative comparisons can be made to statistically differentiate performance.

The most commonly used index for comparing ROC curves is the area  $A$  beneath the curve. This area is equivalent to the probability of success for a diagnostic system identifying both abnormal and normal images in a series of image pairs in which there is always an abnormal and normal image [40]. Examination of Figure 2.11, illustrates how the value for  $A$  effectively varies between 0.5 and 1 and also shows that the farther the ROC curve moves toward the upper left corner, the greater the area under it, and thus the more successful diagnostic system. There is no general



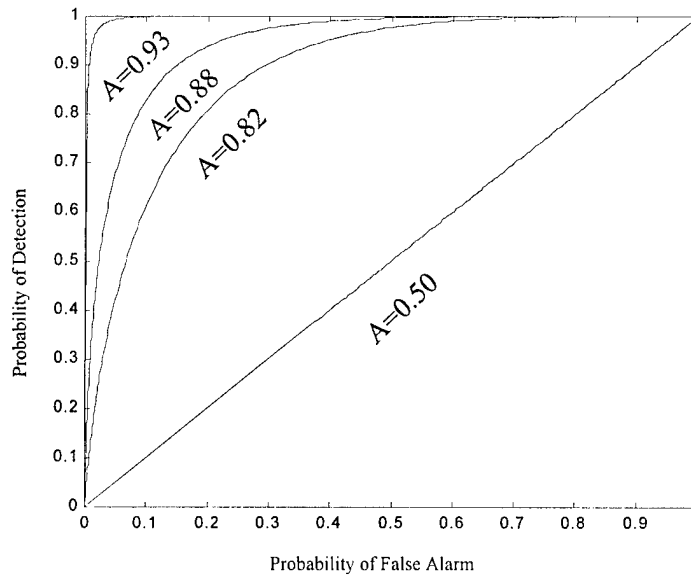


Figure 2.11 Effective range of areas ( $A$ ) under ROC curves.

agreement in the literature with regard to how large an area should be because this is dependent upon the difficulty of the given diagnostic task. However, since  $A = 0.50$  could be achieved by random chance, diagnostic systems seldom have an area below this value [40].

Areas under ROC curves can be obtained in several different ways [50]. If conventional probability axes are used, the ROC points on a curve can be connected with straight lines and the trapezoidal rule can be used to approximate the area. This non-parametric area is designated sometimes in the medical literature as  $P(A)$  and is not typically used since the trapezoidal method greatly underestimates the area when a discrete rating system is used with only a handful of points comprising the ROC curve [49]. This underestimation is due to the way all of the points on the ROC curve are connected with straight lines rather than smooth concave curves. Bradley [20], however, notes that the underestimation of the area (he refers to as AUC) should not be too severe if there are a reasonable number of points comprising the ROC curve. Bradley does not define a reasonable number, but he successfully uses the trapezoidal rule to approximate the area under various ROC curves which have between seven and 15 points [20]. The advantage of using

Table 2.10 Typical discrete rating system used in medical studies.

Classification	Definitely Normal	Probably Normal	Questionable	Probably Abnormal	Definitely Abnormal
Discrete Rating	1	2	3	4	5

Table 2.11 Discrete rating of 109 CT images (data from McNeil and Hanley, 1982 [49]).

Discrete Rating Classification		1	2	3	4	5
Truth	Normal	33	6	6	11	2
	Abnormal	3	2	2	11	33

the trapezoidal approach is that it does not rely on any assumptions regarding the underlying distributions of the target and non-target classes. Also, as will be discussed below, the trapezoidal estimate is exactly the same quantity measured using the Wilcoxon test of ranks.

Since discrete rating systems, as illustrated in Table 2.10 are typically used in medical studies [49], a demonstration will be made here to show how a ROC curve is generated using such a discrete rating system and to describe how a parametric method is commonly used to compute the area under the generated ROC curve. Consider for example, a single reader (i.e., a human expert) assigned with the task of classifying computed tomographic (CT) images obtained from 109 patients with neurological problems [49]. Using the discrete rating system shown in Table 2.10, the reader rates each image, with known disease status, with a discrete rating of 1,2,3,4, or 5 as shown in Table 2.11. In order to generate the ROC curve, the points (pairs of detection and false alarm probabilities) must first be computed. These points are obtained by using different discrete ratings as the decision threshold between normal and abnormal images. For example, images are first classified as abnormal ( $\mathcal{A}$ ) only if they are given a discrete rating of 5 (definitely abnormal) by the reader while the remaining images are classified as normal ( $\mathcal{N}$ ). Using the knowledge of the true class of the images, a standard  $2 \times 2$  confusion matrix can be constructed and the probability of detection and false alarm for this decision threshold represented by the discrete rating of 5 can be computed. Then the decision threshold is changed by now classifying images as abnormal only

Table 2.12  $2 \times 2$  Confusion matrices for varying decision thresholds applied to 109 CT images (data from McNeil and Hanley, 1982 [49]).

Classification for Decision Threshold of:																
Truth		5			4			3			2			1		
			$\mathcal{N}$	$\mathcal{A}$		$\mathcal{N}$	$\mathcal{A}$		$\mathcal{N}$	$\mathcal{A}$		$\mathcal{N}$	$\mathcal{A}$		$\mathcal{N}$	$\mathcal{A}$
		$\mathcal{N}$	56	2	$\mathcal{N}$	45	13	$\mathcal{N}$	38	19	$\mathcal{N}$	33	25	$\mathcal{N}$	0	58
		$\mathcal{A}$	18	33	$\mathcal{A}$	7	44	$\mathcal{A}$	5	46	$\mathcal{A}$	3	48	$\mathcal{A}$	0	51
		$P_D = 0.65$			$P_D = 0.86$			$P_D = 0.90$			$P_D = 0.94$			$P_D = 1.00$		
		$P_{FA} = 0.03$			$P_{FA} = 0.22$			$P_{FA} = 0.33$			$P_{FA} = 0.43$			$P_{FA} = 1.00$		

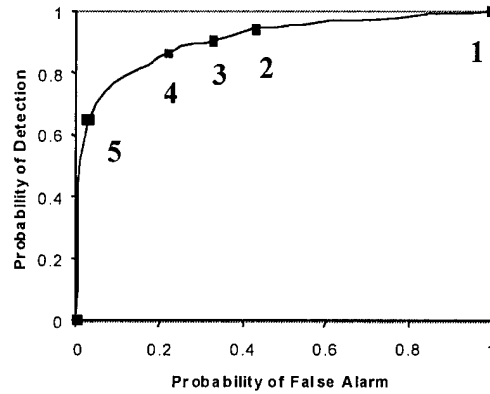


Figure 2.12 ROC curve generated from 109 CT images using the discrete rating system (data from McNeil and Hanley, 1982 [49]).

those images given a discrete rating of 4 or 5 by the reader. The decision threshold is varied again and again until all images are classified as abnormal, i.e., images with discrete ratings ranging from 1 to 5 are classified as abnormal. For each decision threshold setting, the resulting  $2 \times 2$  confusion matrix is constructed and the associated probabilities are computed as shown in Table 2.12.

The ROC curve is then generated by plotting these five probability pairs along with the zero point ( $P_{FA} = 0, P_D = 0$ ) on conventional probability axes as shown in Figure 2.12.

If ROC curves are assumed to be based on underlying Gaussian distributions, then the ROC can be plotted using binormal (normal-deviate) coordinate axes as shown in Figure 2.13. The area  $A$  under the ROC curve can then be computed as

$$A = A_z = \Phi \left( \frac{a}{\sqrt{1 + b^2}} \right) \quad (2.38)$$

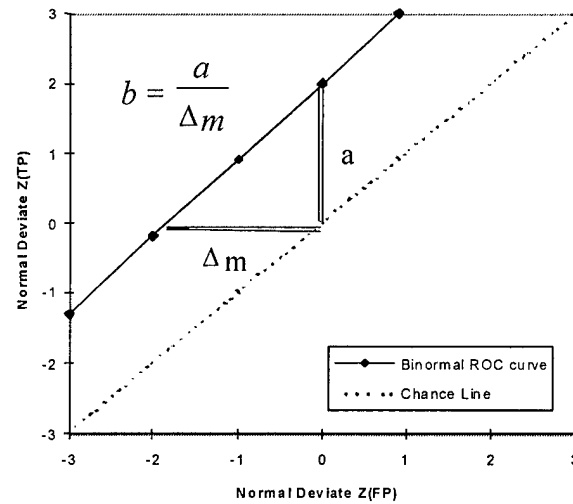


Figure 2.13 ROC curve plotted using normal-deviate coordinates.

where  $\Phi$  is the cumulative standard normal distribution function,  $a$  and  $b$  are the ROC curve's two parameters (the  $y$ -intercept  $a$  and the slope  $b$  as depicted in Figure 2.13 above) in this binormal space, and  $A_z$  denotes that the area has been computed using the binormal assumption. Several computer programs have been developed for ROC analysis that compute fitted curves and calculate areas and their standard errors based on these two parameters. These programs are variations on an original program by Dorfman and Alf [25] which calculates maximum-likelihood estimates and variances of  $a$  and  $b$ .

The area  $A$  and its standard error can also be computed using the Wilcoxon statistic,  $W$ . The results are approximations for a discrete rating system and more exact for a continuous system.  $W$  is usually computed to test whether the levels of some quantitative variable  $x$ , such as the rating, in one population ( $\mathcal{A}$ , for abnormal) tend to be greater than in a second population ( $\mathcal{N}$ , for normal), without making any assumptions about the distributions of  $x$  in the two populations [49]. The null hypothesis is that  $x$  is not a useful discriminator, i.e.,  $\Pr(x_{\mathcal{A}} > x_{\mathcal{N}}) \leq 0.5$ . With a sample of size

$n_{\mathcal{A}}$  from  $\mathcal{A}$  and  $n_{\mathcal{N}}$  from  $\mathcal{N}$ , the Wilcoxon statistic  $W$  is defined by

$$W = \frac{1}{n_{\mathcal{A}} \cdot n_{\mathcal{N}}} \sum_1^{n_{\mathcal{A}}} \sum_1^{n_{\mathcal{N}}} S(x_{\mathcal{A}}, x_{\mathcal{N}}) \quad (2.39)$$

$$\text{where } S(x_{\mathcal{A}}, x_{\mathcal{N}}) = \left\{ \begin{array}{ll} 1 & \text{if } x_{\mathcal{A}} > x_{\mathcal{N}} \\ \frac{1}{2} & \text{if } x_{\mathcal{A}} = x_{\mathcal{N}} \quad (\text{discrete case only}) \\ 0 & \text{if } x_{\mathcal{A}} < x_{\mathcal{N}} \end{array} \right\}$$

such that all  $n_{\mathcal{A}} \cdot n_{\mathcal{N}}$  possible comparisons between the  $n_{\mathcal{A}}$  sample  $x_{\mathcal{A}}$ 's and the  $n_{\mathcal{N}}$  sample  $x_{\mathcal{N}}$ 's are made. For a discrete rating system, the quantity  $W$  can be thought of as an estimate of the *true* area under the curve, i.e., the area one would obtain with an infinite sample and a continuous rating scale. McNeil and Hanley [49] show that this area estimate is exactly the same estimate obtained when using the trapezoidal approximation method discussed above.

The differences between two areas can be tested for statistical significance by comparing the critical ratio  $z$  defined as

$$z = \frac{\text{Area}_1 - \text{Area}_2}{SE(\text{Area}_1 - \text{Area}_2)} \quad (2.40)$$

with the table of the normal distribution [50,64]. In general, the standard error of the area difference ( $SE$ ) can be very complex. Swets and Pickett [64] provide a general expression for  $SE$  to take into account the three types of variances that may be present in a paired data comparison:

1. variance induced by using the selected data set.
2. variance induced by having one system classify the same data set more than once.
3. variance induced by having multiple systems classify the same data set.

Instead of comparing ROC curves using a single performance index, like  $A_z$ , a simultaneous comparison can be made of the two parameters ( $a$  and  $b$ ) which characterize ROC curves in binormal

coordinates [50]. This is a more rigorous statistical comparison approach than the use of area under the ROC curve because while the identity between both curves' parameters can only exist if there is complete coincidence of the curves, two curves may have the same area but not be coincidental curves. Metz et al. [52] provide a test statistic for the differences in the parameters that follows a Chi-square distribution.

Rather than using single or joint accuracy indices that are based on an entire ROC curve and that provide general assessments of system performance, an investigator may desire to use an accuracy index based on one ROC point when the full ROC curve is determined. The most direct and easily interpretable single-parameter index based on one operating point of a fully determined ROC is the value of  $\Pr(TP)$  corresponding to some carefully selected reference  $\Pr(FP)$  [64]. This  $TP$  point index is useful when comparing ROC curves that cross or when investigating for differences in ROC curves in a specific range of interest which may not be detected in any global test. McNeil and Hanley [50] and Metz et al. [52] show how to determine if statistical differences in the  $TP$  point indices exist between two different ROC curves based on the same data set.

As an alternative measure, the area above the ROC curve can be measured by integrating from some lower limit for  $\Pr(TP)$  to 1.0, with better performance minimizing this area. This area is denoted as  $A_{\varphi}^{+}$ , where the  $+$  denotes the area above the ROC curve and  $\varphi$  is the lower limit for  $\Pr(TP)$  [30].

*2.3.4 Multinomial Selection Procedures.* The problem of determining which of  $k$  systems is most likely to be the best performer based on some objective performance measure is known as the multinomial selection problem (MSP). Mathematically the MSP is described as follows [53]. Let  $X_{ji}$  represent the  $i^{th}$  replication from system  $j$  of some performance measure. Each system  $(\pi_j, j = 1, 2, \dots, k)$  has an unknown constant probability  $(p_j, j = 1, 2, \dots, k)$  of having the largest value of the performance measure. The best system is defined as the one most likely to have the largest performance measure in any comparison across all systems. Such a comparison corresponds

to a multinomial trial, where one and only one system can win in any given trial. The objective of the MSP is to find the system, given a limited amount of data, that is most likely to be the best performer in a single trial among the systems, rather than identifying the best average performer over the long run.

Procedure BEM (Bechhofer, Elmaghraby, and Morse [16]) is a classical solution procedure for the MSP. BEM prescribes a minimum number  $\nu^*$  of independent vector replications across all systems such that the probability of correctly selecting the true best system (PCS) meets or exceeds a prespecified probability. On the assumption that larger is better, BEM selects the system having the largest value of the performance measure in more replications than any other, as the best system. The probability of success,  $p_j$ , i.e., the probability of being the best, for each system can be estimated as

$$p_j = \frac{Y_j}{\nu} \quad (2.41)$$

where  $Y_j$  is the number of successes of system  $\pi_j$  for  $\nu$  replications. PCS can be calculated using BEM for a fixed  $k$  and  $\nu$  as

$$PCS^{BEM}([\mathbf{p}]) = \sum_{\mathbf{y}} \frac{1}{t(\mathbf{y})} \frac{\nu!}{\prod_{j=1}^k y_{[j]}!} \prod_{j=1}^k p_{[j]}^{y_{[j]}} \quad (2.42)$$

where the summation is over all vectors  $\mathbf{y} = (y_{[1]}, y_{[2]}, \dots, y_{[k]})$  such that [53] :

1.  $\sum_{j=1}^k y_j = \nu$  ;
2.  $y_{[k]} \geq y_{[j]}$  where  $y_{[k]}$  denotes the ranked number of successes for each system;
3.  $t(\mathbf{y})$  is a function of  $y_{[1]}, y_{[2]}, \dots, y_{[k]}$  representing the number of populations tied for the most wins;
4.  $p_{[j]}$  denotes the ranked success probabilities for each system and  $[\mathbf{p}]$  denotes the set of  $p_{[j]}$ .

Miller, et al. [53] propose an alternative approach, they call Procedure AVC (All Vector Comparisons) designed to obtain a higher PCS by performing all possible comparisons across all systems for a given set of system performance data. The advantage of AVC is that a smaller number of replications are needed to achieve a desired PCS. However, a necessary condition for applying AVC is that the performance measures must be independent of the replication number, which is not the case for detection problems.



### III. A Family of Metrics for Comparing Receiver Operating Characteristic Curves

#### 3.1 Overview

In the pattern recognition community, a commonly held assumption is that for the case of unlimited data, a limiting ROC curve and a classification strategy exists. In reality only finite data are available so only an approximate ROC curve can be constructed. As more (finite) data are added the approximate ROC curve is updated. Performing this process repeatedly yields a sequence of approximate ROC curves. In this chapter a family of metrics for comparing two ROC curves is presented. This family of metrics enables a proof of convergence for these ROC curves. This ROC convergence theorem is important because it provides the basis for a framework for the comparison of ROC curves and hence, the comparison of classifiers.

The research in this chapter summarizes Alsing *et al.* [5, 7]. The chapter is organized as follows. It begins with a definition of a ROC curve given finite data. This is followed by a description of the proposed family of metrics for comparing two ROC curves and the presentation of the theorem for ROC convergence. The proposed metrics are applied to two diagnostic problems to illustrate the usefulness of these metrics, especially when the ROC curves of competing classifiers overlap.

#### 3.2 ROC Curve Given Finite Data

**3.2.1 Mathematical Description of a ROC Curve.** As described in Section 2.2.6, a ROC curve is generated by varying the decision threshold  $\theta$  over all possible values. As  $\theta$  is varied from a low decision threshold value to a high decision threshold value,  $P_{FP}(\theta)$  and  $P_{TP}(\theta)$  both take on values between 0 and 1. Mathematically then,  $P_{FP}(\theta)$  and  $P_{TP}(\theta)$  have the following properties:

1.  $P_{FP}(\theta)$  and  $P_{TP}(\theta)$  are non-decreasing functions of  $\theta$ .
2.  $P_{FP}(\theta)$  and  $P_{TP}(\theta)$  are upper semi-continuous functions of  $\theta$ .

3.  $P_{FP}(\theta)$  and  $P_{TP}(\theta)$  are implicit functions of the random variable  $Z$  (Equations 2.22 and 2.25).

Define the set of possible  $\theta$  values for the random variable  $Z$  as the set  $\Theta$ . For the example shown in Figure 2.9 (page 2-21),  $\Theta = (-\infty, \infty)$ . In general  $\Theta$  can be some subset of  $\Re$ . A *proper* ROC curve starts at  $(P_{FP} = 0, P_{TP} = 0)$  and ends at  $(P_{FP} = 1, P_{TP} = 1)$  [27]. In order to ensure that a *proper* ROC curve is generated the following definition is made.

**Definition III.1.** *The set  $\Theta = (a, b) \subset \Re$  is said to be an **admissible threshold set** for the random variable  $Z$  if*

$$\begin{aligned} \lim_{\theta \rightarrow a^+} P_{FP}(\theta) &= 0 \text{ and } \lim_{\theta \rightarrow a^+} P_{TP}(\theta) = 0 \\ \lim_{\theta \rightarrow b^-} P_{FP}(\theta) &= 1 \text{ and } \lim_{\theta \rightarrow b^-} P_{TP}(\theta) = 1 \end{aligned}$$

Let  $\Theta$  denote an admissible threshold set throughout this dissertation. A ROC Trajectory  $\mathcal{F}$  can then be defined over the admissible threshold set  $\Theta$  as the ordered triple or 3-tuple

$$\mathcal{F} = \{(\theta, P_{FP}(\theta), P_{TP}(\theta)) : \theta \in \Theta\} \quad (3.1)$$

Let  $\mathbf{P}(\theta) \equiv (P_{FP}(\theta), P_{TP}(\theta))$ , then the ROC Trajectory  $\mathcal{F}$  can be defined in the more compact notation as

$$\mathcal{F} = \{(\theta, \mathbf{P}(\theta)) : \theta \in \Theta\} \quad (3.2)$$

A ROC curve  $f$  is then simply the projection of  $\mathcal{F}$  onto the  $(P_{FP}, P_{TP})$  plane

$$f = \{\mathbf{P}(\theta) : \theta \in \Theta\} = \{(P_{FP}(\theta), P_{TP}(\theta)) : \theta \in \Theta\} \quad (3.3)$$

**Remark 1.**  $f$  is a relation. In some cases,  $f$  is a function, i.e., if  $(p, q) \in f$  then, given  $p$  there exists a unique value  $q$ . Thus, we write  $q = f(p)$  to denote this unique value. For example, let  $F_1$  and  $F_2$  denote the cumulative distribution functions of the scalar feature  $z$  for Class 1 and Class 2 data, respectively. Let  $Z_1$  and  $Z_2$  denote random variables with these distributions. When  $Z_i$  is discrete, the ROC function is a set of discrete points. When  $F_1$  and  $F_2$  are continuous, a closed-form expression for the ROC function,  $f$ , can be written [47] as

$$f(p) = 1 - F_2(F_1^{-1}(1 - p)) \quad (3.4)$$

for all  $p \in [0, 1]$ , assuming the converse relation  $F_1^{-1}$  is a function. Lloyd [47] points out that for both the discrete and continuous cases,  $f$  is nothing more than the distribution function of  $1 - F_1(Z_2)$ . Statistically, this is the non-null distribution function of the  $p$ -value [55],  $1 - F_1(Z_2)$ , for testing the null hypothesis that a given feature  $z$  comes from Class 1 [47].

The set of all possible ROC curves for an admissible set  $\Theta$  will be denoted by  $R = R(\Theta)$ .

That is

$$R = \{f : \exists Z \text{ and } \Theta \text{ which is admissible for } Z\}. \quad (3.5)$$

**3.2.2 Empirical ROC Curves from Unknown Data Distributions.** In practice the distribution functions are not known. Three approaches for constructing ROC curves are available:

1. Assume the form of the distributions (binormal most common in medical diagnostic research employing the *rating* method) to fit ROC curves [49, 64].
2. Estimate the unknown distributions from the sample data [38] and use Lloyd's [47] ROC function  $f$  (Equation 3.4).

3. Approximate ROC curves by using Equation 3.3 above with the estimated probability of false positive,  $\hat{P}_{FP}$ , and the estimated probability of true positive,  $\hat{P}_{TP}$  [43, 44].

Consider approach 3 which is commonly used in ATR [43, 44]. The estimated probabilities used in approach 3 are random variables because they both depend upon the actual finite data used [2, 3]. As an illustration, consider again a two-class problem, but with multiple variables or a feature vector  $\mathbf{x}$ . Let  $z \in \Theta$  here be the real-valued output (in particular,  $\Theta = [0, 1]$ ) of a pattern recognition algorithm, representing the probability for class 1 (non-target) membership, i.e.,  $z = \Pr(c = C_1 | \mathbf{x})$ . For classifiers based on linear or quadratic discriminant functions,  $z$  can be taken as the *a posterior* probability when weighted against the alternatives [45], using Bayes Rule for classification with equal costs for misclassification and equal prior probabilities assumed:

A feature vector  $\mathbf{x}$  is assigned to class  $c = C_k$  if  $P(C_k | \mathbf{x}) > P(C_j | \mathbf{x})$  for all  $j \neq k$  [18].

In the case of a multi-layer perceptron artificial neural network conditions exist [18, 59, 62] where  $z$  estimates, in the limit, the *a posterior* probability as well.

Let  $\mathbf{x}_i$ ,  $i = 1, \dots, 2n$ , where  $2n$  is the total number of feature vectors, be the finite test data  $\mathcal{D}^{(n)} = \{\mathbf{x}_i \in \mathbb{R}^v : i = 1, \dots, 2n\}$  used, where  $v$  is the number of variables or features and assume for now that the number of data points for each class in any test data set are equal to  $n$ . Let  $\omega$  be the specific instantiation of this finite *vector* data that can be drawn from the set of all possible data where the event space or set  $\Omega$  is the set of all possible instantiations of this finite *vector* data. For every feature vector  $\mathbf{x}$ , a particular classifier will generate a scalar output  $z$ , resulting in the output space  $\mathcal{Z}^{(n)} = \{z_i \in [0, 1] : i = 1, \dots, 2n\}$ . Given an integer  $n$ , a decision threshold  $\theta \in \Theta$ , and a given instantiation  $\omega \in \Omega$  of the data, the estimated probability of false positive,  $\hat{P}_{FP}^{(n)}$ , and

the estimated probability of true positive,  $\hat{P}_{TP}^{(n)}$ , are

$$\hat{P}_{FP}^{(n)}(\omega, \theta) = \frac{\text{card}\{z_i < \theta \mid C_1, i = 1, \dots, n\}}{n} = \frac{\sum_{i=1}^n \chi_{[0, \theta]}(z_i | C_1)}{n} \quad (3.6)$$

$$\hat{P}_{TP}^{(n)}(\omega, \theta) = \frac{\text{card}\{z_i < \theta \mid C_2, i = 1, \dots, n\}}{n} = \frac{\sum_{i=1}^n \chi_{[0, \theta]}(z_i | C_2)}{n}, \quad (3.7)$$

where  $\text{card}\{\mathcal{E}\}$  is the cardinality of event  $\mathcal{E}$  or the number of times event  $\mathcal{E}$  occurs and  $\chi_S(z_i | C_k) \equiv \chi_{S \times C_k}[(z_i, c_i)]$  is the characteristic function defined by

$$\chi_{S \times C_k}[(z_i, c_i)] = \begin{cases} 1, & \text{if } z_i \in S \text{ and } c_i \in C_k \\ 0, & \text{otherwise} \end{cases}. \quad (3.8)$$

For example,

$$\begin{aligned} \chi_{[0, \theta]}(z_i | C_1) &= \chi_{[0, \theta] \times C_1}[(z_i, c_i)] \\ &= \begin{cases} 1, & \text{if } z_i \in [0, \theta] \text{ and } c_i \in C_1 \\ 0, & \text{otherwise} \end{cases}. \end{aligned} \quad (3.9)$$

The estimated or empirical ROC curve,  $\hat{f}^{(n)}(\omega)$ , is defined by varying the decision threshold,  $\theta$ , over its entire range,  $\Theta$ , specifically

$$\hat{f}^{(n)}(\omega) = \left\{ \left( \hat{P}_{FP}^{(n)}(\omega, \theta), \hat{P}_{TP}^{(n)}(\omega, \theta) \right) : \theta \in \Theta \right\} \quad (3.10)$$

$$= \left\{ \hat{\mathbf{P}}^{(n)}(\omega, \theta) : \theta \in \Theta \right\}. \quad (3.11)$$

In order to make meaningful comparisons between empirical ROC curves, as  $n$  becomes large, in the sense that more feature vectors  $\mathbf{x}$  are added to the data,  $\hat{f}^{(n)}(\omega)$  must converges in probability

to a limiting ROC curve  $f \in \mathcal{R}(\Theta)$ , defined as

$$f = \{(P_{FP}(\theta), P_{TP}(\theta)) : \theta \in \Theta\} = \{\mathbf{P}(\theta) : \theta \in \Theta\}. \quad (3.12)$$

If a limiting ROC curve does not exist, then the comparisons of empirical ROC curves are not valid. A family of metrics is proposed in the following section that enable a proof of convergence for these curves.

### 3.3 Comparison of ROC Curves

**3.3.1 Definition of metric and metric spaces.** The distances between two ROC curves is determined with a metric. Bartle and Sherbert provide the following definition [13].

**Definition III.2 (metric).** *A metric on a non-empty set  $S$  is a function  $d : S \times S \rightarrow \mathbb{R}$  that satisfies the following properties:*

1.  $d(x, y) \geq 0$  for all  $x, y \in S$  (positivity);
2.  $d(x, y) = 0$  if and only if  $x = y$  (definiteness);
3.  $d(x, y) = d(y, x)$  for all  $x, y \in S$  (symmetry);
4.  $d(x, y) \leq d(x, z) + d(z, y)$  for all  $x, y, z \in S$  (triangle inequality).

If property 1, 3 and 4 hold, but property 2 does not hold true, then the function  $d$  is said to be a pseudo-metric on  $S$ .

**Definition III.3 (metric space).** *A metric space  $(S, d)$  is a non-empty set  $S$  with a metric  $d$  defined on  $S$ .*

**Definition III.4 (pseudo-metric space).** *A pseudo-metric space  $(S, d)$  is a non-empty set  $S$  with a pseudo-metric  $d$  defined on  $S$ .*

Some examples of metric spaces are  $S = \mathbb{R}^2 = \{\vec{x} = (x_1, x_2) | x_i \in \mathbb{R}\}$  and the metric  $\rho_q$

$$\rho_q(\vec{x}, \vec{y}) = (|x_1 - y_1|^q + |x_2 - y_2|^q)^{\frac{1}{q}}. \quad (3.13)$$

for each  $1 \leq q < \infty$ . The metric  $\rho_1$  is known as the Manhattan metric while  $\rho_2$  is known as the Euclidean metric. For  $q = \infty$ , define the metric  $\rho_\infty$  to be

$$\rho_\infty(\vec{x}, \vec{y}) = \max \{ |x_1 - y_1|, |x_2 - y_2| \}, \quad (3.14)$$

which is known as the infinity metric.

**Definition III.5 (equivalent metrics).** Assuming  $(S, d_\alpha)$  and  $(S, d_\beta)$  are metric spaces, the metrics  $d_\alpha$  and  $d_\beta$  are said to be equivalent metrics on  $S$  if there exist constants  $k, K > 0$  such that

$$kd_\beta(x, y) \leq d_\alpha(x, y) \leq Kd_\beta(x, y) \text{ for all } x, y \in S. \quad (3.15)$$

**Theorem III.1 (equivalent metrics on reals).** All metrics on  $\mathbb{R}^2$  are equivalent.

The proof of this theorem is straight forward using Definition III.5, see Naylor and Sell [56].

**3.3.2 Area Under the ROC Curve (AUC).** Typically in the literature, ROC curves are compared using the area under the ROC curve (AUC) as a performance measure [20, 27, 38, 49, 64]. The classifier with the largest AUC is then considered overall better than its competitors. In this section AUC is shown to be an unsuitable quantifier because the difference in AUCs is not a metric.

Let  $f$  and  $g$  be two ROC curves. Define the delta area,  $\delta$ , as the difference of the areas under the ROC curves, that is

**Definition III.6 (Delta area).** Let  $f, g \in \mathcal{R}$ , then the difference in the area is defined as

$$\delta(f, g) = \left| \int_0^1 f(p)dp - \int_0^1 g(p)dp \right|. \quad (3.16)$$

**Theorem III.2.**  $\delta$  is a pseudo-metric on  $\mathcal{R}(\Theta)$  and  $(\mathcal{R}(\Theta), \delta)$  is a pseudo-metric space.

The proof of this theorem is easily accomplished by showing that  $\delta$  satisfies all the properties of a metric except for the definiteness property. Since  $\delta$  is a pseudo-metric and not a metric,  $\delta$  cannot be used to prove ROC convergence. Even if  $\delta(f^{(n)}, f) \rightarrow 0$  as  $n \rightarrow \infty$ ,  $f^{(n)}(p)$  may not converge to  $f(p)$  for every  $p \in [0, 1]$ . Consider the two ROC curves  $f$  and  $g$ ,

$$f(p) = \frac{1}{2}(2p-1)^3 + \frac{1}{2} \quad (3.17)$$

$$g(p) = p \quad (3.18)$$

shown in Figure 3.1. Both  $f$  and  $g$  have exactly the same AUC and hence  $\delta = 0$ , but they are clearly not the same curve. The implication is that the AUC or the difference in areas may not always be suitable for comparing ROC curves. Rather than using ROC areas, a family of suitable metrics are proposed below for enabling a proof of ROC convergence.

**3.3.3 Definition of proposed ROC metrics.** Let  $1 \leq r < \infty$  and let  $\rho$  be any metric on  $\mathbb{R}^2$ . Given two ROC curves  $f, g \in \mathcal{R}(\Theta)$ ,

$$f = \left\{ \left( P_{FP}^{(f)}(\theta), P_{TP}^{(f)}(\theta) \right) : \theta \in \Theta \right\} = \left\{ \mathbf{P}^{(f)}(\theta) : \theta \in \Theta \right\} \quad (3.19)$$

$$g = \left\{ \left( P_{FP}^{(g)}(\theta), P_{TP}^{(g)}(\theta) \right) : \theta \in \Theta \right\} = \left\{ \mathbf{P}^{(g)}(\theta) : \theta \in \Theta \right\}, \quad (3.20)$$



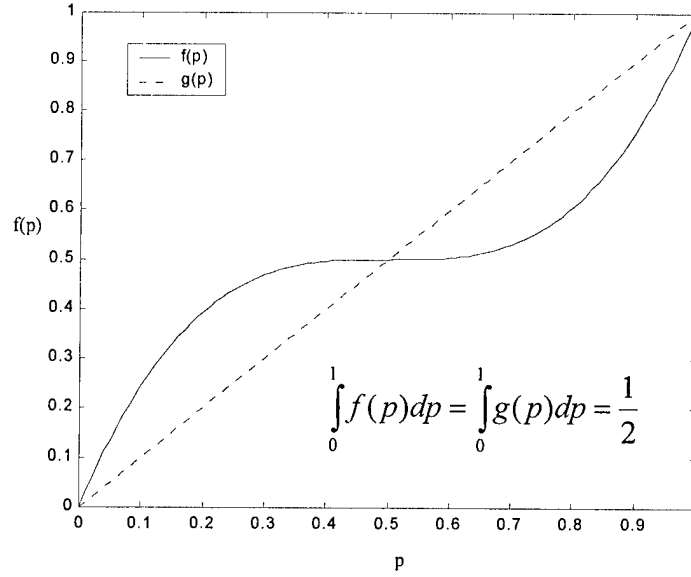


Figure 3.1 Both  $f$  and  $g$  have exactly the same area, but they are clearly not the same function.

define the mapping  $d_{\rho,r}$  as the family

$$d_{\rho,r}(f, g) = \left( \int_{\Theta} \rho \left( \mathbf{P}^{(f)}(\theta), \mathbf{P}^{(g)}(\theta) \right)^r d\theta \right)^{\frac{1}{r}}. \quad (3.21)$$

**Theorem III.3.** For every  $r \in [1, \infty]$ ,  $(R(\Theta), d_{\rho,r})$  is a metric space.

### Proof of Theorem III.3

Let  $1 \leq r < \infty$  and let  $\rho$  be any metric on  $\mathbb{R}^2$ . Given two ROC curves  $f, g \in R(\Theta)$ , it is sufficient to prove that  $d_{\rho,r}(f, g)$  satisfies the four required properties of a metric (Definition III.2, page 3-6).

1. For each  $\theta \in \Theta$ ,  $\rho(\mathbf{P}^{(f)}(\theta), \mathbf{P}^{(g)}(\theta)) \geq 0$  since  $\rho$  is a metric. This implies  $\left( \int_{\Theta} \rho(\mathbf{P}^{(f)}(\theta), \mathbf{P}^{(g)}(\theta))^r d\theta \right)^{\frac{1}{r}} \geq 0$ .
0. Therefore,  $d_{\rho,r}(f, g) \geq 0$  and  $d_{\rho,r}$  satisfies the positivity property.
2. Definiteness proof.

(a) (*only if part*) Let  $d_{\rho,r}(f, g) = 0$ . Then  $\left( \int_{\Theta} \rho(\mathbf{P}^{(f)}(\theta), \mathbf{P}^{(g)}(\theta))^r d\theta \right)^{\frac{1}{r}} = 0$  which implies  $\rho(\mathbf{P}^{(f)}(\theta), \mathbf{P}^{(g)}(\theta)) = 0$  for each  $\theta \in \Theta$ . Since  $\rho$  is a metric, then  $\mathbf{P}^{(f)}(\theta) = \mathbf{P}^{(g)}(\theta)$ . But  $\mathbf{P}^{(f)}(\theta) = \mathbf{P}^{(g)}(\theta)$  for each  $\theta \in \Theta$  which implies  $f = g$ .

(b) (*if part*) Assume  $f = g$ . Then  $\mathbf{P}^{(f)}(\theta) = \mathbf{P}^{(g)}(\theta)$  for all  $\theta \in \Theta$ . And  $\rho(\mathbf{P}^{(f)}(\theta), \mathbf{P}^{(g)}(\theta)) = 0$  for all  $\theta \in \Theta$  since  $\rho$  is a metric. This implies  $\left( \int_{\Theta} \rho(\mathbf{P}^{(f)}(\theta), \mathbf{P}^{(g)}(\theta))^r d\theta \right)^{\frac{1}{r}} = 0$  which implies  $d_{\rho,r}(f, g) = 0$ . Therefore,  $d_{\rho,r}$  satisfies the definiteness property.

3.  $d_{\rho,r}(f, g) = \left( \int_{\Theta} \rho(\mathbf{P}^{(f)}(\theta), \mathbf{P}^{(g)}(\theta))^r d\theta \right)^{\frac{1}{r}} = \left( \int_{\Theta} \rho(\mathbf{P}^{(g)}(\theta), \mathbf{P}^{(f)}(\theta))^r d\theta \right)^{\frac{1}{r}} = d_{\rho,r}(g, f)$  since  $\rho$  is a metric. Therefore,  $d_{\rho,r}$  satisfies the symmetry property.

4. Let  $f, g, h \in \mathcal{R}(\Theta)$ . For each  $\theta \in \Theta$ ,  $\rho(\mathbf{P}^{(f)}(\theta), \mathbf{P}^{(g)}(\theta)) \leq \rho(\mathbf{P}^{(f)}(\theta), \mathbf{P}^{(h)}(\theta)) + \rho(\mathbf{P}^{(h)}(\theta), \mathbf{P}^{(g)}(\theta))$  since  $\rho$  is a metric. Therefore, for  $1 \leq r < \infty$

$$\begin{aligned} d_{\rho,r}(f, g) &= \left( \int_{\Theta} \rho(\mathbf{P}^{(f)}(\theta), \mathbf{P}^{(g)}(\theta))^r d\theta \right)^{\frac{1}{r}} \\ &\leq \left( \int_{\Theta} \rho(\mathbf{P}^{(f)}(\theta), \mathbf{P}^{(h)}(\theta))^r d\theta \right)^{\frac{1}{r}} + \left( \int_{\Theta} \rho(\mathbf{P}^{(h)}(\theta), \mathbf{P}^{(g)}(\theta))^r d\theta \right)^{\frac{1}{r}} \\ &= d_{\rho,r}(f, h) + d_{\rho,r}(h, g) \end{aligned}$$

by r-norm properties of functions [56]. Therefore,  $d_{\rho,r}(f, g) \leq d_{\rho,r}(f, h) + d_{\rho,r}(h, g)$  and  $d_{\rho,r}$  satisfies the triangle inequality.

**3.3.4 Definition of Empirical ROC Curves.** For a given instantiation  $\omega \in \Omega$  of the data, let  $\hat{f}(\omega) \in \mathcal{R}(\Theta)$ . Let  $\mathcal{RR}(\Omega, \Theta)$  denote the set of estimated or empirical ROC curves  $\hat{f}(\omega)$ , that is

$$\begin{aligned} \mathcal{RR}(\Omega, \Theta) &= \left\{ \hat{f}(\omega) : \omega \in \Omega \right\} \\ &= \left\{ \left\{ \left( \hat{P}_{FP}(\omega, \theta), \hat{P}_{TP}(\omega, \theta) \right) : \theta \in \Theta \right\} : \omega \in \Omega \right\}. \end{aligned} \tag{3.22}$$

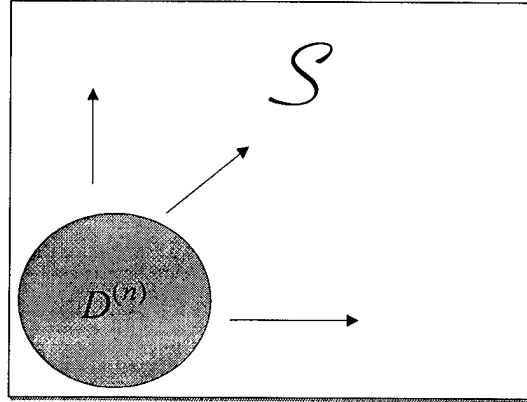


Figure 3.2 Pictorial representation of the convergence in the Hausdorff metric of the set  $\mathcal{D}^{(n)}$  for  $n$  data points to the set  $S$  of all data points.

**3.3.5 ROC Convergence Theorem.** Let  $S \subset \mathbb{R}^v$ , where  $v$  is the number of variables or features, be the set from which feature vectors  $\mathbf{x}$  are drawn. Let  $\mathcal{D}^{(n)} \subset S$  be the set of feature vectors  $\mathbf{x}$  for finite  $n$ , i.e.,  $\mathcal{D}^{(n)} = \{\mathbf{x}_i \in S : i = 1, \dots, 2n\}$ . As  $n$  grows (more data are collected) assume that  $\mathcal{D}^{(n)}$  converges to  $S$ . Under these conditions, the type of convergence assumed is a sequence of sets converging in the Hausdorff metric,  $d_H$  [12]. This type of convergence of sets requires that  $\mathcal{D}^{(n)}$  grows to span  $S$  as more data are collected rather than just become a small subset of  $S$  (Figure 3.2). Let  $\hat{f}^{(n)}(\omega) \in \mathbb{R}(\Omega, \Theta)$  be a sequence of empirical ROC curves. For large  $n$ , assuming that  $\mathcal{D}^{(n)}$  approaches  $S$ , then  $\hat{f}^{(n)}(\omega)$  becomes a better estimate of  $f \in \mathbb{R}(\Theta)$ . This convergence is stated rigorously with the following theorem.

**Theorem III.4 (ROC Convergence).** *If  $\{\mathcal{D}^{(n)}\}$  converges to  $S$ , i.e., given  $\varepsilon > 0$ , there exists  $N$  such that for all  $n > N$ ,  $d_H(\mathcal{D}^{(n)}, S) < \varepsilon$ , then  $\{\hat{f}^{(n)}(\omega)\}$  converges to  $f$ , i.e., given  $\varepsilon > 0$ , there exists  $N$  such that for all  $n > N$ ,  $\Pr\left(\left\{\omega \in \Omega : d_{\rho,r}(\hat{f}^{(n)}(\omega), f) \geq \varepsilon\right\}\right) < \varepsilon$ .*

The proof of this theorem is given in Appendix A.

### 3.4 Application of Proposed ROC Metric

In this section a demonstration is given using one of the proposed ROC metrics, namely  $d_{\rho_1,1}$ ,

$$d_{\rho_1,1}(f, g) = \int_{\Theta} \rho_1(\mathbf{P}^{(f)}(\theta), \mathbf{P}^{(g)}(\theta)) d\theta \quad (3.23)$$

along with the commonly used performance measure, AUC [20, 64]. Since AUC is measured on a scale from 0 to 1, an average metric distance is defined as

$$\text{avg metric distance} = \frac{d_{\rho_1,1}(f, g)}{\mu(\Theta)} = \frac{\int_{\Theta} \rho_1(\mathbf{P}^{(f)}(\theta), \mathbf{P}^{(g)}(\theta)) d\theta}{\mu(\Theta)}. \quad (3.24)$$

where  $\mu$  is a measure on  $\Re$  (possibly Lebesgue) such that  $0 < \mu(\Theta) < \infty$ . Choose  $\Delta \in \Re$  such that  $(\frac{1}{\Delta} + 1)$  is an integer. Then the ROC curves for each classifier are generated by using  $(\frac{1}{\Delta} + 1)$  discrete thresholds  $\theta_i$  and the admissible set  $\Theta$  is then given by

$$\Theta = \{\theta_1, \theta_2, \dots, \theta_i, \theta_{i+1}, \dots, \theta_m : \theta_1 = 0, \theta_{i+1} = \theta_i + \Delta, \theta_m = 1\}. \quad (3.25)$$

For a given  $\Delta$  and defining  $\mu$  such that  $\mu(\Theta) = m$ , the average metric distance can be approximated as follows:

$$\text{avg metric distance} \approx \frac{\sum_{i=1}^m \rho_1(\mathbf{P}^{(f)}(\theta_i), \mathbf{P}^{(g)}(\theta_i))}{m}. \quad (3.26)$$

Using this approximation, the scale for this average metric distance, like the scale for AUC, extends from 0 to 1. However, an average metric distance of 0 implies no difference between ROC curves  $f$  and  $g$ , while an average metric distance of 1 implies maximum difference between ROC curves  $f$  and  $g$ . Both this average metric distance and AUC are applied to a real-world application—the University of Wisconsin Breast Cancer Diagnosis problem. The data set for this application is available from the University of California-Irvine [66].

Table 3.1 Feature rankings for University of Wisconsin Breast Cancer Diagnosis Data Set obtained by using the signal-to-noise ratio algorithm [14]. Rankings averaged over 30 different test data sets.

Feature #	1	2	3	4	5	6	7	8	9
Feature Name	thick	size	shape	adhesion	epithelial	bare nuclei	bland chromatin	norm nuclei	mitoses
Mean Feature Ranking	2.87	5.43	5.03	5.70	7.40	1.80	5.50	5.53	5.73
Standard Error	0.22	0.47	0.52	0.42	0.30	0.23	0.40	0.36	0.38

*3.4.1 Data Description.* The University of Wisconsin Breast Cancer Diagnosis Data Set consists of 699 patterns of which 458 are benign samples and 241 are malignant samples. Each of these patterns consists of nine measurements taken from fine needle aspirates from a patient's breast. These measurements consisted of (1) clump thickness, (2) uniformity of cell size, (3) uniformity of cell shape, (4) marginal adhesion, (5) single epithelial cell size, (6) bare nuclei, (7) bland chromatin, (8) normal nucleoli, and (9) mitoses. All nine measurements were graded on an integer scale from one to ten, with one being the closest to benign and ten being the most malignant. Sixteen samples of feature number 6, bare nuclei, were missing from the data set. Rather than dropping this feature and risk losing an important feature, the missing values are estimated using a linear regression with feature 6 as the independent variable and the other features as the dependent variables. Predictions of feature 6 are then used to estimate the missing values.

*3.4.2 Experiment #1.* The goal of the first experiment is to compare three different classifiers for which the relative ordering is already known. To obtain these classifiers the feature ranking results (Table 3.1) of the signal-to-noise ratio (SNR) feature screening method as applied to the University of Wisconsin Breast Cancer Diagnosis problem [14] is used. The top two ranked features (bare nuclei and clump thickness) are used for classifier 1; the two features (bland chromatin and normal nucleoli) ranked in the middle for classifier 2; and the bottom two ranked features (mitoses and single epithelial cell size) for classifier 3. Thirty artificial neural networks are trained for each classifier using 210 randomly selected samples for training, 140 samples for internal

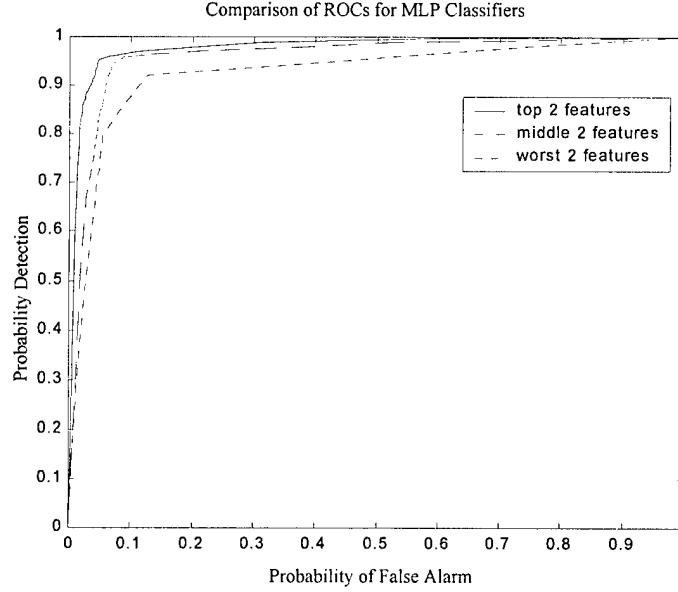


Figure 3.3 Experiment 1: average ROC curves for three MLP Classifiers. ROC Curves averaged over 30 different test data sets.

validation, and 349 for independent testing. All neural networks are multi-layer perceptrons (MLPs) trained using MATLAB's adaptive learning algorithm (TRAINGDX) with an initial learning rate of 0.01 [23]. This algorithm also employs momentum with a momentum constant of 0.9. All features are standardized to zero mean and unit variance. One hidden layer is employed with 18 nodes. All activation functions are sigmoidal.

ROC curves for each MLP classifier are generated using 101 discrete decision thresholds ( $\Delta = 0.01$ ) for each test data set. The average ROC curves over 30 different test data sets for the three MLP classifiers are shown in Figure 3.3. The ROC curves appear as expected with the ROC for classifier 1 closer to the “northwest corner” of the graph (perfection:  $P_{FP} = 0$ ,  $P_{TP} = 1$ ) followed by classifier 2 and then classifier 3. The area under each ROC curve for the three classifiers (Table 3.2, 2nd column) is estimated by using the trapezoidal method for all thirty ROC curves for each classifier and computing the mean. The mean areas agree with the expected ordering of the classifiers. The average metric distance is computed in a similar way to compare each classifier

to the  $\theta = P_{TP} = P_{FP}$  diagonal. The average metric distances (Table 3.2, 3rd column) also agree with the expected ordering of the classifiers.

The areas can be converted to delta areas by computing the absolute differences in areas between the ROC curves for each pair of classifiers. These delta areas are shown in a distance matrix (Table 3.3, lower triangular matrix), where 0 in a cell matrix implies that the two classifiers corresponding to the row and column number of the respective cell have exactly the same area. Table 3.3, upper triangular matrix shows the average metric distances between each pair of ROC curves. Both upper and lower triangular matrices display similar patterns and agree with the expected ordering of the classifiers.

*3.4.3 Experiment #2.* The goal of the second experiment is to compare three different types of classifiers which have ROC curves that overlap. A multi-layer perceptron (MLP) artificial neural network, a linear statistical classifier, and a quadratic statistical classifier are used. The top two ranked features (bare nuclei and clump thickness) are used to construct all three classifiers. The MLP neural network is exactly the same as described above. The linear statistical classifier is employed using a discriminant analysis function using equal (pooled) covariance matrices for each class (benign and malignant) while the quadratic statistical classifier employs unequal covariance matrices for each class. Thirty statistical classifiers are trained for each type using 350 randomly selected samples for training, and 349 for independent testing.

ROC curves for each classifier are generated using 101 discrete decision thresholds ( $\Delta = 0.01$ ) for each test data set. The average ROC curves over 30 different test data sets for the three different types of classifiers are shown in Figure 3.4 along with the respective areas in Table 3.4 (2nd column) and the corresponding delta areas in the lower triangular matrix in Table 3.5. Because the ROC curves overlap, there is some difficulty in determining which classifier is overall the best. The areas (Table 3.4, 2nd column) or the delta areas (Table 3.5, lower triangular matrix) do not help eliminate the confusion. The areas for all three classifiers are within 0.007 and the delta area

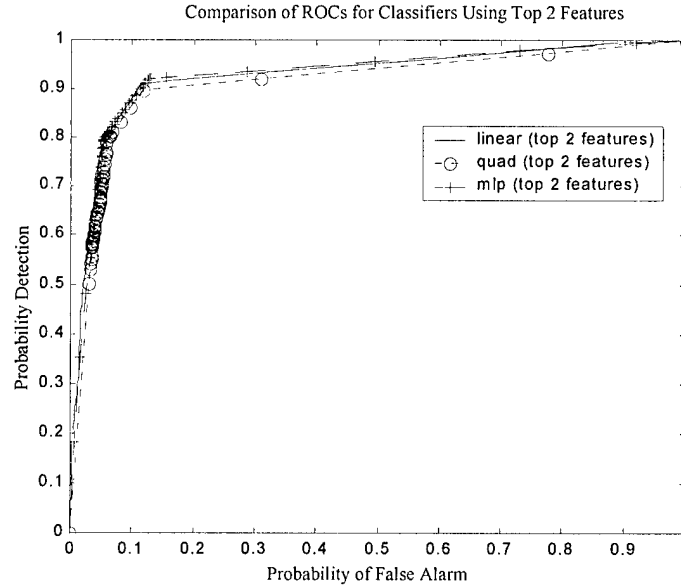


Figure 3.4 Experiment 2: average ROC Curves for linear, quadratic, and MLP classifiers. ROC curves averaged over 30 different test data sets.

between the linear and MLP classifiers is statistically zero. The AUC confidence intervals for the linear and MLP classifiers overlap, while the AUC confidence intervals for the quadratic and the MLP classifiers overlap. Therefore any clear distinction between the three classifiers using AUC is impossible to ascertain.

The average metric distances between each pair of ROC curves using this metric are shown in the upper triangular matrix in Table 3.5. The differences between all three classifiers are easier to ascertain using the average metric distances in the upper triangular matrix in Table 3.5, rather than using the delta areas shown in the lower triangular matrix. Since the average metric distances between each pair of ROC curves is statistically non-zero, the implication is that there are in fact differences in the three curves.

The average metric distances between the ROC curve for each classifier and the  $\theta = P_{TP} = P_{FP}$  diagonal are shown in Table 3.4 (3rd column). The average metric distances from the diagonal distinguish the quadratic classifier as the best of the three (confidence intervals do not overlap).



Table 3.2 Comparison of area under ROC curves and average metric distances from diagonal line for MLP ROCs. Bonferroni confidence intervals based on 30 different test data sets (normal assumption with  $\alpha_{total} = 0.05$ ).

Classifier	Area under ROC curve	Average metric distance from diagonal line
1	$0.9805 \pm 0.0039$	$0.8007 \pm 0.0099$
2	$0.9673 \pm 0.0037$	$0.7496 \pm 0.0188$
3	$0.9319 \pm 0.0045$	$0.6593 \pm 0.0150$

Table 3.3 Distance matrix showing absolute area differences (lower triangular matrix) and average metric distances (upper triangular matrix) between MLP ROC curves. Bonferroni confidence intervals based on 30 different test data sets (normal assumption with  $\alpha_{total} = 0.05$ ).

	Classifier 1	Classifier 2	Classifier 3
Classifier 1	0	$0.1034 \pm 0.0149$	$0.1653 \pm 0.0175$
Classifier 2	$0.0147 \pm 0.0039$	0	$0.1552 \pm 0.0192$
Classifier 3	$0.0486 \pm 0.0057$	$0.0354 \pm 0.0049$	0

Table 3.4 Comparison of area under ROC curves and average metric distances from diagonal line for linear, quadratic, and MLP classifiers. Bonferroni confidence intervals based on 30 different test data sets (normal assumption with  $\alpha_{total} = 0.05$ ).

Classifier	Area under ROC curve	Average metric distance from diagonal line
Linear	$0.9822 \pm 0.0023$	$0.7628 \pm 0.0048$
Quadratic	$0.9755 \pm 0.0032$	$0.8395 \pm 0.0068$
MLP	$0.9805 \pm 0.0039$	$0.8007 \pm 0.0099$

Table 3.5 Distance matrix showing absolute area differences (lower triangular matrix) and average metric distances (upper triangular matrix) between linear, quadratic, and MLP classifiers. Bonferroni confidence intervals based on 30 different test data sets (normal assumption with  $\alpha_{total} = 0.05$ ).

	Linear	Quadratic	MLP
Linear	0	$0.1169 \pm 0.0081$	$0.1101 \pm 0.0105$
Quadratic	$0.0067 \pm 0.0013$	0	$0.0798 \pm 0.0083$
MLP	$0.0025 \pm 0.0027$	$0.0065 \pm 0.0017$	0

An examination of the average ROC curves (Figure 3.4) provides an explanation. The highest concentration of points that comprise the ROC curve for the quadratic classifier occur in a smaller region compared to the majority of points that comprise the ROC curves for the linear and MLP classifiers. This high concentration of points in a smaller region for the quadratic classifier reflects the robustness of the quadratic classifier's performance for various decision thresholds. Since the metric distance is the average taxi-cab distance from each ROC point  $(P_{FP}^{(f)}(\theta_i), P_{TP}^{(f)}(\theta_i))$  on ROC curve  $f$  to its corresponding point  $(\theta_i, \theta_i)$  on the diagonal line  $g$ , it is not surprising that the metric distance for the quadratic statistical classifier is statistically the largest. The choice of the quadratic classifier using the average metric distance from the diagonal then, represents a choice for consistency in performance.

### 3.5 Conclusion

This chapter introduces a family of metrics  $d_{\rho,r}$  for comparing two ROC curves that enables a proof of convergence for these curves. This ROC convergence theorem is important because it provides the basis for a framework for the comparison of ROC curves. The typical ROC performance measure, AUC, used for comparing ROC curves is shown to be an unsuitable metric because the comparison of areas is in fact a pseudo-metric. The experiment comparing MLPs constructed using the salient and non-salient features of the University of Wisconsin Breast Cancer Diagnosis Data Set showed that a particular metric,  $d_{\rho_1,1}$ , from the proposed family of metrics yielded expected classifier rankings that are consistent with AUC. Furthermore, the second experiment comparing linear, quadratic, and MLP classifiers showed that this particular metric provides more insight about classifier differences when the ROC curves for the classifiers overlap. The results of these experiments bolster confidence in the proposed family of metrics presented here as a useful tool in distinguishing between the ROC curves of competing classifiers.

## *IV. Using a Multinomial Selection Procedure in Classifier Evaluation*

### *4.1 Overview*

Multinomial selection procedures have not been previously applied to the detection problem in the pattern recognition literature. This chapter explores the use of a multinomial selection procedure as an alternative to the ROC type analyses discussed above for evaluating competing classifiers and to serve as a baseline for comparing methods.

This chapter is organized in the following manner. In Section 4.2 the multinomial selection procedure as applied to the detection problem is illustrated using the classical two-dimensional exclusive-OR problem. Classification accuracy, ROC analysis, as well as multinomial selection procedure results for two other more difficult discrimination problems are described in Sections 4.3-4.4. Finally, in Section 4.5 the strengths of the multinomial selection procedure are summarized.

### *4.2 Illustration of Multinomial Selection Procedure on XOR Problem*

In order to illustrate how a multinomial selection procedure can be applied to the detection problem, consider first the classical two-dimensional exclusive-OR problem, also known as XOR [18] shown in Figure 4.1. XOR data are randomly generated with 1000 Class 1 and 1000 Class 2 data points. The data are classified using three different types of classifiers: a linear statistical classifier, a quadratic statistical classifier, and a multi-layer perceptron (MLP) artificial neural network. All three classifiers are trained on fifty percent of the data (balanced between the two classes) and tested and compared on the remaining fifty percent. The linear statistical classifier is employed using a discriminant analysis function using equal (pooled) covariance matrices for each class while the quadratic statistical classifier employs unequal covariance matrices for each class. The MLP is trained using MATLAB's adaptive learning algorithm (TRAINGDX) with an initial learning rate of 0.01 [23]. One hidden layer is employed with eight nodes. Forty percent of the training data is used for internal validation of the MLP to prevent over training. The resulting confusion matrices

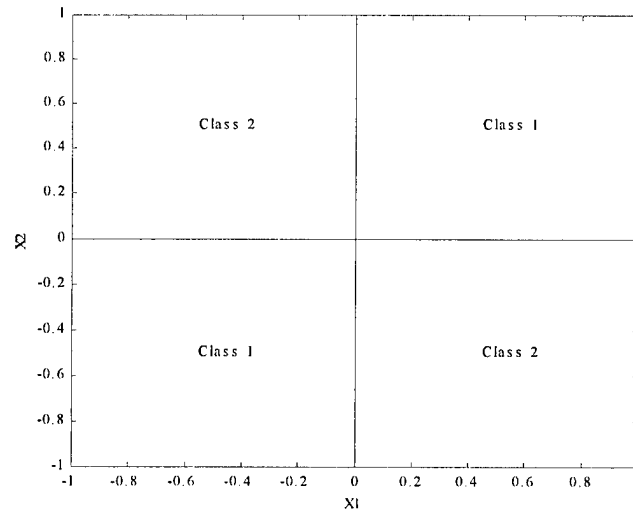


Figure 4.1 The exclusive-OR problem, also known as XOR, consists of data belonging to one of two classes  $C_1$  and  $C_2$  which are not linearly separable.

and classification errors on the test data set for the three classifiers are shown in Tables 4.1 - 4.3. The ROC curves (where Class 2 is considered the target) for the three classifiers for the test data are shown in Figure 4.2, along with the corresponding AUCs in Table 4.4.

As expected, the linear statistical classifier is not much better than a coin toss. However, the MLP appears to have only slightly better performance than the quadratic statistical classifier after examining the classification errors, ROC curves, and corresponding AUCs on the test data set. Instead of comparing the classifiers over the entire test data set at once, a multinomial selection procedure compares the performance of each classifier on each data point using some scoring measure. A logical choice for classifier scores is the estimated class conditional posterior probabilities generated by each classifier for each data point. The test data is first separated into the two independent classes. The BEM multinomial selection procedure is then applied to each class as follows:

1. Given  $\nu_j$  Class  $j$  test data points, compare estimated posterior Class  $j$  probabilities for each classifier.

Table 4.1 Linear classifier confusion matrix for XOR data.

		Linear Classifier	
		Classified As (Reported)	
Actual (Truth)		Class 1	Class 2
	Class 1	259	241
	Class 2	245	255

error = 48.6%

Table 4.2 Quadratic classifier confusion matrix for XOR data.

		Quadratic Classifier	
		Classified As (Reported)	
Actual (Truth)		Class 1	Class 2
	Class 1	483	17
	Class 2	15	485

error = 3.2%

Table 4.3 MLP classifier confusion matrix for XOR data.

		MLP Classifier	
		Classified As (Reported)	
Actual (Truth)		Class 1	Class 2
	Class 1	490	10
	Class 2	12	488

error = 2.2%

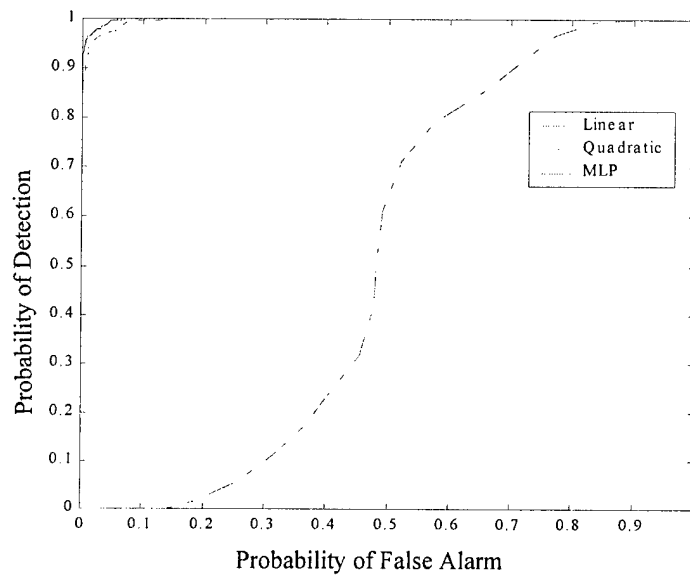


Figure 4.2 ROC curves for XOR problem.

2. Select the best classifier for each data point as the classifier with the maximum estimated posterior Class  $j$  probability.
3. Compute the number of wins/successes  $Y_{i|j}$  for each classifier  $i$  given Class  $j$  data.
4. Let  $Y_{(1)} \leq Y_{(2)} \leq Y_{(3)}$  be the ranked number of successes from Step 3. Select the classifier associated with the largest count,  $Y_{(3)}$ , as the best for Class  $j$ .

This BEM procedure is illustrated in Table 4.5 for Class 1 data and Table 4.6 for Class 2 data. Since the MLP clearly has the most successes for both classes, the conclusion, according to the BEM procedure, is that the MLP is the best of the three systems. However, even with only one test data set, the BEM procedure provides additional information about the competing systems. If the number of successes  $Y_{i|j}$  for each classifier  $i$  given  $\nu_j$  Class  $j$  test data points, is modeled as a single multinomial distribution, a point estimate can be computed for the conditional probability  $P(C_i|X_j)$  of each classifier  $C_i$  being the best given the class  $X_j$  using

$$P(C_i|X_j) = \frac{Y_{i|j}}{\nu_j} \quad (4.1)$$

**Remark 2.** *An inherent assumption in a multinomial distribution is a constant probability of success over all test trials, or in this case all test points. For this type of application of the BEM procedure, it is not altogether clear that the probability of success, i.e., probability of being the best (Equation 4.1) is constant from trial to trial. However, an argument could be made that the trials are still random, and the probabilities of success obtained are still estimates of the probabilities of winning in any randomly selected trial. Such an argument is made to justify the use of Equation 4.1 to generate point estimates for the conditional probabilities of each classifier being the best given the class.*

The point estimates for the probability of being the best classifier along with their corresponding Bonferroni confidence intervals (using Equation 2.14 with  $\alpha_{total} = 0.05$ ) for comparing

Table 4.4 AUCs (area under ROC curve) for XOR data.

Classifier	AUC
Linear	0.5114
Quadratic	0.9970
MLP	0.9988

Table 4.5 BEM procedure illustrated for Class 1 XOR data.

Test Data	Posterior Probabilities			Win/Successes = 1		
Number	Linear	Quad	MLP	Linear	Quad	MLP
1	0.451	0.908	1.000	0	0	1
2	0.395	1.000	0.998	0	1	0
3	0.434	0.988	0.996	0	0	1
4	0.524	0.594	0.994	0	0	1
5	0.589	0.999	1.000	0	0	1
6	0.421	0.606	0.900	0	0	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
498	0.526	0.644	1.000	0	0	1
499	0.454	0.836	1.000	0	0	1
500	0.577	0.979	1.000	0	0	1
	Successes ( $Y_{i 1}$ ) =			Linear 5	Quad 25	MLP 470

Table 4.6 BEM procedure illustrated for Class 2 XOR data.

Test Data	Posterior Probabilities			Win/Successes = 1		
Number	Linear	Quad	MLP	Linear	Quad	MLP
501	0.558	0.998	0.984	0	1	0
502	0.472	0.658	1.000	0	0	1
503	0.550	0.989	0.991	0	0	1
504	0.499	0.958	1.000	0	0	1
505	0.506	0.516	0.844	0	0	1
506	0.448	0.973	1.000	0	0	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
998	0.499	0.938	1.000	0	0	1
999	0.498	0.957	1.000	0	0	1
1000	0.497	0.772	1.000	0	0	1
	Successes ( $Y_{i 2}$ ) =			Linear 4	Quad 69	MLP 427

Table 4.7 Estimates and Bonferroni confidence intervals ( $\alpha_{total} = 0.05$ ) for the probability of being the best classifier given Class 1 data

$P_{best}$	Classifier 1 Linear	Classifier 2 Quadratic	Classifier 3 MLP
$P(C_i X_1)$	0.01	0.05	0.94
CI	[0 0.02]	[0.03 0.05]	[0.91 0.97]

Table 4.8 Estimates and Bonferroni confidence intervals ( $\alpha_{total} = 0.05$ ) for the probability of being the best classifier given Class 2 data

$P_{best}$	Classifier 1 Linear	Classifier 2 Quadratic	Classifier 3 MLP
$P(C_i X_2)$	0.01	0.14	0.85
CI	[0 0.02]	[0.10 0.18]	[0.81 0.89]

the three classifiers are given in Tables 4.7 - 4.8. Since the confidence interval for the MLP does not overlap the confidence intervals for the other two classifiers for either class data, the BEM procedure suggests that the MLP is statistically the best system for both classes. The total probability that each classifier is the best according to the estimated posterior probabilities can be computed using the law of total probability

$$P(C_i) = P(C_i|X_1)P(X_1) + P(C_i|X_2)P(X_2) \quad (4.2)$$

where  $P(X_j)$  are the prior probabilities for each class (for this problem,  $P(X_1) = P(X_2) = 0.5$ ). These total probabilities and their corresponding Bonferroni confidence intervals (Table 4.9) indicate that the MLP is statistically the best classifier for this problem.

The BEM procedure also provides an equation (Equation 2.42, pg. 2-37) for the probability of correct selection  $PCS^{BEM}$ . In order to get a lower bound on  $PCS^{BEM}$ , a least favorable condition

Table 4.9 Estimates and Bonferroni confidence intervals ( $\alpha_{total} = 0.05$ ) for the total probability of being the best classifier for XOR data.

$P_{best}$ (total)	Classifier 1 Linear	Classifier 2 Quadratic	Classifier 3 MLP
$P(C_i)$	0.01	0.09	0.90
CI	[0 0.02]	[0.07 0.12]	[0.88 0.92]



(LFC) is chosen using

$$p_{[1]} = p_{[2]} = \frac{P(C_1|X_j) + P(C_2|X_j)}{2} \quad (4.3)$$

and  $p_{[3]} = P(C_3|X_j)$  for the ranked probabilities of being the best system for each class. Using these probabilities along with the number of test points  $\nu_j$  in each class, and computer code developed by Goldsman [16], the probability of correct selection for each class is estimated to be

$$PCS^{BEM} = 1. \quad (4.4)$$

Because the probability of correct selection is one, the number of actual test data points needed to differentiate between the three classifiers could have been reduced. Using the same ratio  $\hat{\theta}$  between the best and the next best system

$$\hat{\theta} = \frac{p_{[3]}}{\frac{p_{[1]} + p_{[2]}}{2}}. \quad (4.5)$$

Goldsman's code [16], which constructs the tables of Bechhofer, Elmaghraby, and Morse [16], provide the number of test points (Table 4.10) required to achieve  $PCS^{BEM} = 1.0$ . Table 4.10 implies that only 15 test points are needed to differentiate the MLP classifier as being the best classifier for either class.

#### 4.3 Block C Problem

The XOR problem is fairly easy for both the quadratic statistical classifier as well as the MLP. A more challenging discrimination problem is shown in Figure 4.3. Block C data are randomly generated with 1000 Class 1 and 1000 Class 2 data points. The data are again classified as described in Section 4.2 using three different types of classifiers: a linear statistical classifier, a quadratic statistical classifier, and a multi-layer perceptron (MLP) artificial neural network. The

Table 4.10 Values for the Probability of Correct Selection (*PCS*) for various number of test data points for the XOR problem.

Number of Test Points	Class 1 <i>PCS</i>	Class 2 <i>PCS</i>
1	0.9399	0.8540
2	0.9399	0.8540
3	0.9913	0.9514
4	0.9944	0.9656
5	0.9987	0.9838
6	0.9994	0.9905
7	0.9998	0.9952
8	0.9999	0.9971
9	1.0000	0.9985
10	1.0000	0.9992
11	1.0000	0.9995
12	1.0000	0.9997
13	1.0000	0.9999
14	1.0000	0.9999
15	1.0000	1.0000

resulting confusion matrices and classification errors on the test data set for the three classifiers are shown in Tables 4.11 - 4.13. The ROC curves (where Class 2 is again considered the target) for the three classifiers for the test data are shown in Figure 4.4, along with the corresponding AUCs in Table 4.14. As expected again, the linear statistical classifier is not much better than a coin toss for this problem. While both the quadratic statistical classifier and the MLP show better performance than the linear classifier, according to the classification errors and the AUCs, the MLP appears to have better performance. The dominance of the MLP is clearly seen in Table 4.15 where the point estimates and Bonferroni confidence intervals (using Equation 2.14 with  $\alpha_{total} = 0.05$ ) for comparing the three classifiers are computed as shown in Section 4.2 using the BEM multinomial selection procedure and the law of total probability (Equation 4.2). Using Goldsman's code as in the previous section [16], the probability of correctly selecting the MLP as the best classifier for each class is estimated to be

$$PCS^{BEM} = 1. \quad (4.6)$$

Table 4.11 Linear classifier confusion matrix for Block C data.

		Linear Classifier	
		Classified As (Reported)	
Actual (Truth)		Class 1	Class 2
	Class 1	263	237
	Class 2	238	262

error = 47.5%

Table 4.12 Quadratic classifier confusion matrix for Block C data.

		Quadratic Classifier	
		Classified As (Reported)	
Actual (Truth)		Class 1	Class 2
	Class 1	438	62
	Class 2	37	463

error = 9.9%

Table 4.13 MLP classifier confusion matrix for Block C data.

		MLP Classifier	
		Classified As (Reported)	
Actual (Truth)		Class 1	Class 2
	Class 1	477	23
	Class 2	4	496

error = 2.7%

Table 4.14 AUCs (area under ROC curve) for Block C data.

Classifier	AUC
Linear	0.5471
Quadratic	0.9520
MLP	0.9864

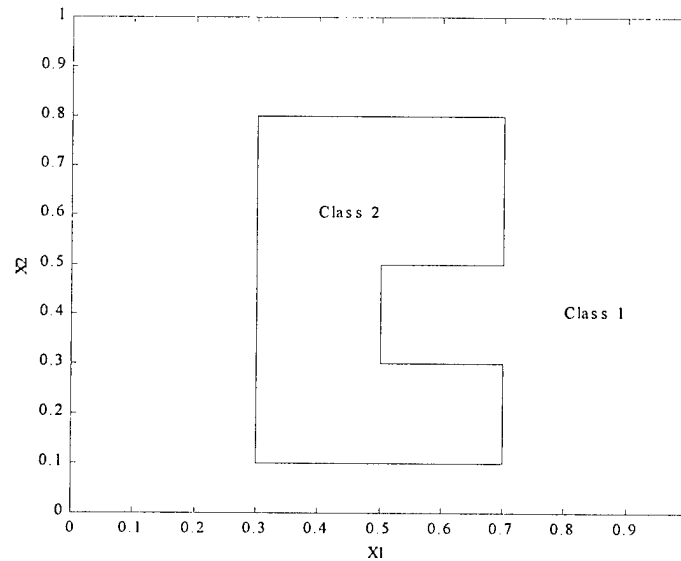


Figure 4.3 The Block C problem consists of data belonging to one of two classes  $C_1$  and  $C_2$  which are not linearly separable (taken from Belue, 1995 [17]).

Table 4.15 Estimates and Bonferroni confidence intervals ( $\alpha_{total} = 0.05$ ) for the total probability of being the best classifier for Block C data.

$P_{best}$ (total)	Classifier 1 Linear	Classifier 2 Quadratic	Classifier 3 MLP
$P(C_i)$	0.02	0.04	0.94
CI	[0.01 0.03]	[0.02 0.06]	[0.92 0.96]

giving confidence to the choice of the MLP. Also, this same code can be used to determine that only 12 test points are needed to differentiate the MLP classifier as being the best classifier for either class.

#### 4.4 Iron Cross Problem

An even more challenging discrimination problem than either the XOR or the Block C problem is shown in Figure 4.5. Iron Cross data are randomly generated with 1000 Class 1 and 1000 Class 2 data points. The data are again classified as described in Section 4.2 using three different types of classifiers: a linear statistical classifier, a quadratic statistical classifier, and a multi-layer

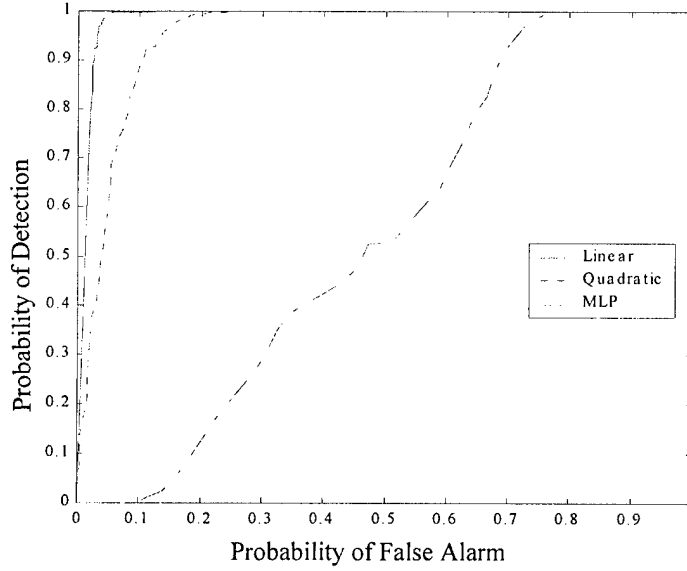


Figure 4.4 ROC curves for Block C data.

perceptron (MLP) artificial neural network (for this problem the MLP required 12 hidden nodes). The resulting confusion matrices and classification errors on the test data set for the three classifiers are shown in Tables 4.16 - 4.18. The ROC curves (where Class 2 is again considered the target) for the three classifiers for the test data are shown in Figure 4.6, along with the corresponding AUCs in Table 4.19. As expected again, the linear statistical classifier is not much better than a coin toss for this problem. However, the MLP clearly shows better performance than both the linear and quadratic classifier, according to the classification errors and the AUCs. The dominance of the MLP is also clearly seen in Table 4.20 where the point estimates and Bonferroni confidence intervals (using Equation 2.14 with  $\alpha_{total} = 0.05$ ) for comparing the three classifiers are computed as shown in Section 4.2 using the BEM multinomial selection procedure and the law of total probability (Equation 4.2). Using Goldsman's code as in the previous section [16], the probability of correctly selecting the MLP as the best classifier for each class is estimated to be

$$PCS^{BEM} = 1. \quad (4.7)$$

Table 4.16 Linear classifier confusion matrix for Iron Cross data.

		Linear Classifier	
		Classified As (Reported)	
Actual (Truth)		Class 1	Class 2
	Class 1	250	250
	Class 2	236	264

error = 48.6%

Table 4.17 Quadratic classifier confusion matrix for Iron Cross data.

		Quadratic Classifier	
		Classified As (Reported)	
Actual (Truth)		Class 1	Class 2
	Class 1	291	209
	Class 2	185	315

error = 39.4%

Table 4.18 MLP classifier confusion matrix for Iron Cross data.

		MLP Classifier	
		Classified As (Reported)	
Actual (Truth)		Class 1	Class 2
	Class 1	472	28
	Class 2	10	490

error = 3.8%

Table 4.19 AUCs (area under ROC curve) for Iron Cross data.

Classifier	AUC
Linear	0.5117
Quadratic	0.6526
MLP	0.9970

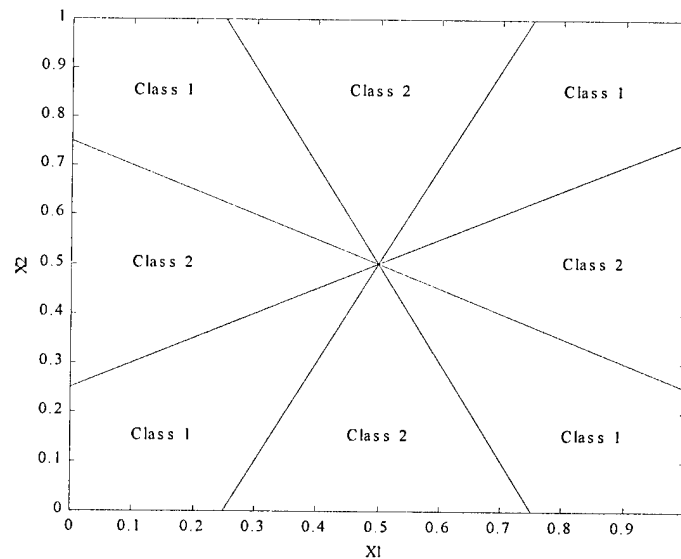


Figure 4.5 A very challenging discrimination problem, termed the Iron Cross problem.

Table 4.20 Estimates and Bonferroni confidence intervals ( $\alpha_{total} = 0.05$ ) for the total probability of being the best classifier for Iron Cross data.

$P_{best}$ (total)	Classifier 1 Linear	Classifier 2 Quadratic	Classifier 3 MLP
$P(C_i)$	0.02	0.02	0.96
CI	[0.01 0.03]	[0.01 0.03]	[0.94 0.98]

giving confidence to the choice of the MLP. Also, this same code can be used to determine that only 9 test points are needed to differentiate the MLP classifier as being the best classifier for either class.

#### 4.5 Conclusions

This chapter introduces a multinomial selection procedure as an alternative to ROC analysis for evaluating competing classifiers. Three discrimination problems of varying difficulty are used to illustrate the method. For the XOR problem both classification accuracy and the area under the ROC curve do not clearly distinguish between the quadratic statistical classifier and the multilayer perceptron (MLP) classifier, while the probability of being the best classifier from the multinomial

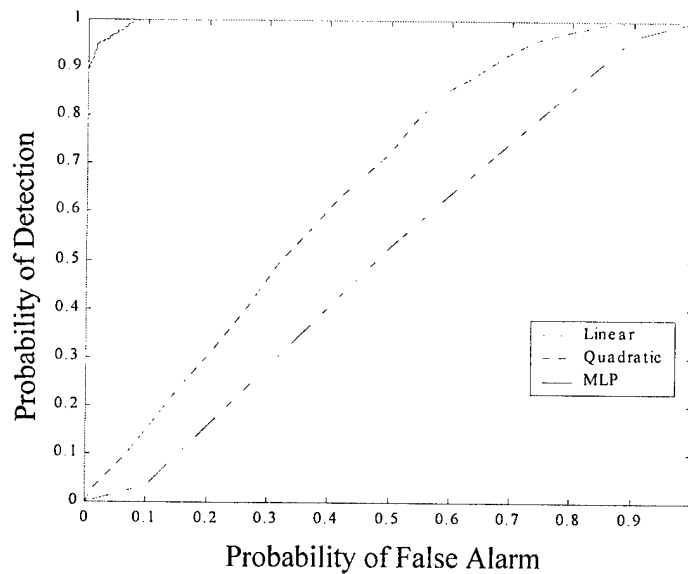


Figure 4.6 ROC curves for Iron Cross data.

selection procedure does. For all three discrimination problems, the ordering of the classifiers according to the multinomial selection procedure agrees with the ordering of classifiers based on classification accuracy and ROC analysis using AUCs. However, in all three problems the dominance of the MLP classifier is more easily ascertained using the multinomial selection procedure, which requires only one pair of training and testing data to estimate a performance measure with confidence intervals. Also, the multinomial selection procedure provides additional information about the number of test data points needed to distinguish between the classifiers. These results provide confidence in the multinomial selection procedure as a useful tool in distinguishing between competing classifiers.



## *V. ATR Application*

### *5.1 Overview*

This chapter provides comparisons between the methodologies introduced in Chapter III and IV and typical approaches on an ATR application. As discussed in Section 1.2.1, the USAF is especially interested in objectively evaluating algorithm upgrades to their ATR system, MSTAR. The Air Force Research Laboratory Sensors Directorate manages the MSTAR program and is leading the effort toward promoting generally sound evaluation practices in ATR research [67]. The recent public release of high resolution SAR data by the MSTAR program has provided a unique opportunity to promote and evaluate SAR ATR algorithm development. The application summarized in this chapter uses two statistical classifiers and an ANN classifier on this SAR data to illustrate the methodologies developed in this dissertation.

The research in this chapter has its foundation in two referee reviewed papers [4, 19]. This chapter is organized as follows. Section 5.2 provides a description of the SAR data used. Section 5.3 describes the experimental setup and classifiers employed. Section 5.4 provides the results obtained using the various methodologies to compare the competing classifiers. Discussion of the results are given in Section 5.5 and conclusions are provided in Section 5.6.

### *5.2 Data Description*

The SAR data used consist of 1-D high range resolution (HRR) data taken from the MSTAR Public Data Set [67]. The HRR data was formed by processing X-band  $1 \times 1$  foot resolution complex spotlight SAR images for ten target types collected over full  $360^\circ$  aspect coverage at  $15^\circ$  and  $17^\circ$  depression angles as part of the MSTAR program data collections 1 and 2, scene 1. Two target types (BMP2 armored personnel carrier and T72 tank) have additional configuration variants yielding a total of 22 targets in the data set. The targets are listed in Table 5.1 by MSTAR class, type, and

Table 5.1 Target listing for MSTAR Public Data Set.

MSTAR Class	Target Type	Serial Number	MSTAR Data Product Source
Armored Personnel Carrier	BMP2	C21 9563 9566	MSTAR (Public) Targets
Tank	T72	132 812 S7 A04 A05 A07 A10 A32 A62 A63 A64	MSTAR (Public) Targets MSTAR (Public) Targets MSTAR (Public) Targets MSTAR/IU (Public) T72 Variants MSTAR/IU (Public) T72 Variants MSTAR/IU (Public) T72 Variants MSTAR/IU (Public) T72 Variants MSTAR/IU (Public) T72 Variants MSTAR/IU (Public) T72 Variants MSTAR/IU (Public) T72 Variants MSTAR/IU (Public) T72 Variants
Tank	T62	A51	MSTAR/IU (Public) Mixed Targets
Gun	ZSU23-4 2S1	D08 B01	MSTAR/IU (Public) Mixed Targets
Transport	BTR70 BTR60	C71 K10YT7532	MSTAR (Public) Targets MSTAR/IU (Public) Mixed Targets
Truck	BRDM2 ZIL131	E71 E12	MSTAR/IU (Public) Mixed Targets
Bulldozer	D7	92V13015	MSTAR/IU (Public) Mixed Targets

serial number along with the public source of the SAR chips. The public CDs can be requested via the World Wide Web at <http://www.mbvlab.wpafb.af.mil/public/MBVDATA> [67].

For this application the average HRR profile is used for each aspect angle that data is available. Although the images for the targets are spaced in  $1^\circ$  increments over full aspect, the aspect sampling is not uniform and results in data dropouts. The HRR profile used contains the average of the center eight signatures in the range/angle data matrix after each signature range bin is magnitude detected, normalized by the mean signature power and power transformed (exponent = 0.2) [67]. Figure 5.1 shows an example of an average HRR profile for a BMP2 (serial # C21) armored personnel carrier at an aspect angle of  $150.1914^\circ$ .

Standard approaches for HRR identification include the use of the entire range profile as the feature vector or the selection of peak amplitudes within a specified range bin [54]. For this application, the peak amplitudes within range bin numbers 21-30, 31-40, 41-50, 51-60, 61-70, and

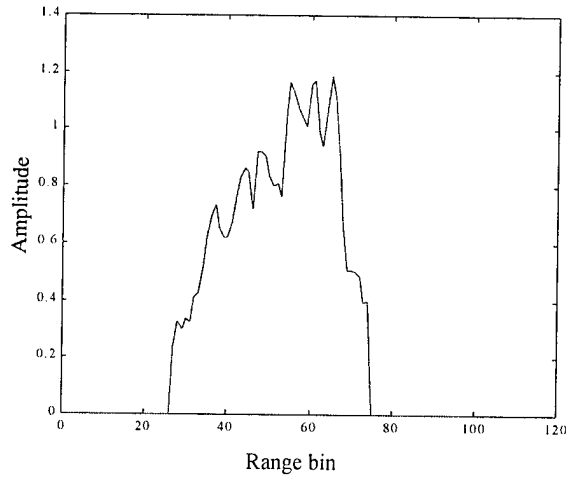


Figure 5.1 Average HRR profile for a BMP2 (serial # C21) armored personnel carrier. Depression angle is  $17^\circ$  and aspect angle is  $150.1914^\circ$ .

71-80 are used to characterize the average HRR profiles [4]. These six amplitudes together with their corresponding aspect angle comprise the feature vector used for HRR classification.

### 5.3 Experiment Description

**5.3.1 Experimental setup.** The experiment classifies targets into two classes: non-targets or “confusers” (class 1) and targets specified for attack (class 2) [67]. The confusers consist of the two trucks (BRDM2 and ZIL131) and the bulldozer from the MSTAR Public Data Set (Table 5.1). The targets of interest consist of the BMP2 armored personnel carrier and the T72 tank. Following the evaluation guidelines set forth by Velten, et al. [67],  $17^\circ$  depression angle data is used for training and  $15^\circ$  depression angle data is used for testing. Furthermore, only the BMP2 armored personnel carrier, serial # C21, and the T72 tank, serial #132, are used for training while all the variants of the BMP2 armored personnel carrier and T72 tank shown in Table 5.1 are used for testing. This setup provides for training at a specific extended operating point (EOC), i.e., nominal BMP2 armored personnel carrier and nominal T72 tank at  $17^\circ$  depression angle, and testing at various EOCs, i.e., variants of the BMP2 armored personnel carrier and T72 tank at

15° depression angle [4]. The testing data was divided into 33 trials, each consisting of the three confusers and one combination of the variants of the BMP2 armored personnel carrier and T72 tank.

*5.3.2 Classifiers.* A linear statistical classifier, a quadratic statistical classifier, and an MLP ANN are used. The linear statistical classifier is employed using a discriminant analysis function using equal (pooled) covariance matrices for each class while the quadratic statistical classifier employs unequal covariance matrices for each class. Both statistical classifiers are trained on the one training data set and then tested on the 33 different testing data sets in order to measure the robustness of the classifiers. Thirty-three different neural networks are trained on the same training data set, but using a different random initialization for the weights. These ANNs are then tested on the 33 different testing data sets for robustness. All neural networks are MLPs trained using MATLAB's adaptive learning algorithm (TRAINGDX) with an initial learning rate of 0.01 [23]. This algorithm also employs momentum with a momentum constant of 0.9. All features are standardized to zero mean and unit variance. One hidden layer is employed with 12 nodes. All activation functions are sigmoidal.

#### 5.4 Results

The confusion matrices for all three classifiers are shown in Tables 5.2-5.4. These confusion matrices represent classifier performance at the Bayes optimal point, i.e., classification performance at a decision threshold that minimizes total classification error. The raw numbers in the confusion matrix are the sums over all 33 test data sets. The percentages are the means of 33 probabilities conditioned on the rows for each test data set. Bonferroni confidence intervals for comparing the two independent probability estimators in each confusion matrix for each classifier are computed using the normal assumption with  $\alpha_{total} = 0.05$ .

Table 5.2 Linear classifier confusion matrix for ATR application. Raw numbers are the sums over 33 test data sets. Percentages and corresponding half-lengths of the Bonferroni confidence intervals are also based on 33 independent test data sets (normal assumption with  $\alpha_{total} = 0.05$ ).

Linear Classifier CLASSIFICATION		
	Confuser	Target
Confuser	12375 (57.60 $\pm$ 0.00%)	9108 (42.40 $\pm$ 0.00%)
Target	125 (0.82 $\pm$ 0.23%)	14645 (99.18 $\pm$ 0.23%)

Table 5.3 Quadratic classifier confusion matrix for ATR application. Raw numbers are the sums over 33 test data sets. Percentages and corresponding half-lengths of the Bonferroni confidence intervals are also based on 33 independent test data sets (normal assumption with  $\alpha_{total} = 0.05$ ).

Quadratic Classifier CLASSIFICATION		
	Confuser	Target
Confuser	18843 (87.71 $\pm$ 0.00%)	2640 (12.29 $\pm$ 0.00%)
Target	4962 (33.21 $\pm$ 2.30%)	9808 (66.79 $\pm$ 2.30%)

Table 5.4 MLP classifier confusion matrix for ATR application. Raw numbers are the sums over 33 test data sets. Percentages and corresponding half-lengths of the confidence intervals are also based on 33 independent test data sets (normal assumption with  $\alpha = 0.05$ ).

MLP Classifier CLASSIFICATION		
	Confuser	Target
Confuser	20401 (94.96 $\pm$ 0.67%)	1082 (5.04 $\pm$ 0.67%)
Target	4370 (29.53 $\pm$ 1.38%)	10400 (70.47 $\pm$ 1.38%)

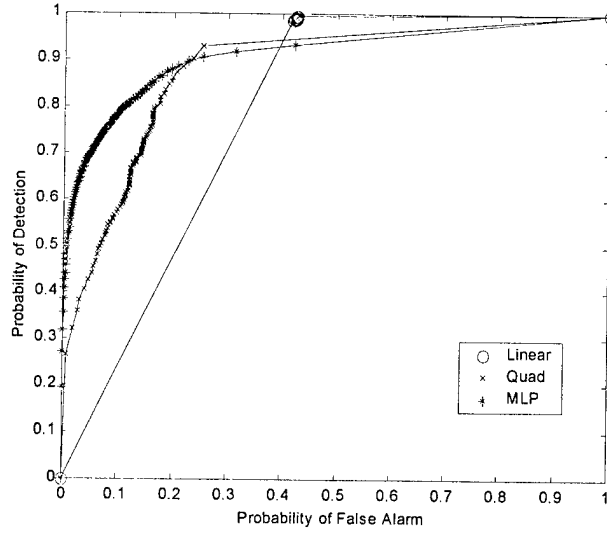


Figure 5.2 Average ROC curves for ATR application. Curves averaged over 33 independent test data sets.

ROC curves for each classifier are generated by using 101 discrete decision thresholds for each test data set. The average (over 33 test data sets) ROC curves are shown in Figure 5.2. The area under the ROC curve (AUC) for the three classifiers is estimated by using the trapezoidal method for all 33 ROC curves for each classifier and computing the mean. The average (over all thresholds) of the proposed ROC metric  $d_{\rho_1,1}$  (Section 3.4) is computed for all 33 ROC curves to compare each classifier to the  $\theta = P_{TP} = P_{FP}$  diagonal (chance) line for each test data set. The mean metric distances along with the mean AUCs are shown in Table 5.5.

The BEM multinomial selection procedure as described in Section 4.2 is also applied to the ATR problem. The conditional probabilities for each classifier being the best given the class (confuser or target) are computed for each test data set. The prior probability of confusers (or targets) is then estimated for each test data set by computing the ratio of the number of confusers (or targets) to the total number of confusers and targets. Using these estimates for the prior probabilities, the total probability that each classifier is the best for each test data set is

Table 5.5 Summary of various performance measures for ATR application. Mean estimates and half-lengths for simultaneous Bonferroni confidence intervals based on 33 independent test data sets (normal assumption with  $\alpha_{total} = 0.05$ ).

Classifier	CA	AUC	Metric Distance	$P_{best}(total)$
Linear	$0.7451 \pm 0.0032$	$0.7866 \pm 0.0013$	$0.7286 \pm 0.0026$	$0.7251 \pm 0.0034$
Quadratic	$0.7910 \pm 0.0111$	$0.8833 \pm 0.0070$	$0.5524 \pm 0.0140$	$0.1376 \pm 0.0072$
MLP	$0.8500 \pm 0.0061$	$0.9146 \pm 0.0089$	$0.6155 \pm 0.0098$	$0.1373 \pm 0.0067$

computed using the law of total probability (Equation 4.2). The mean (over 33 test data sets) total probabilities are also given in Table 5.5.

Table 5.5 also contains estimates for classification accuracy which is the typical performance index used in pattern recognition. Classification accuracy (CA) is estimated for a specific decision threshold, namely the Bayes optimal point. The Bayes optimal point is the decision threshold for which the total misclassification error is a minimum. Classification accuracy is defined for this application as follows

$$CA = \frac{\text{number of confusers and targets classified correctly}}{\text{total number of confusers and targets}}. \quad (5.1)$$

Classification accuracy is computed for each test data set for each classifier. The mean (over 33 test data sets) CAs are reported in Table 5.5.

For all the performance measures shown in Table 5.5, the half-lengths for confidence intervals are also reported. These half-lengths represent simultaneous 95% Bonferroni [46] confidence intervals for comparing the three classifiers.

## 5.5 Discussion

Using classification accuracy (CA) and AUC as the performance measures, Table 5.5 indicates that the MLP is the best classifier. However, a closer examination of the average ROC curves

Table 5.6 Comparison of the trapezoidal approximation (AUC), the binormal approximation ( $A_z$ ), and the Wilcoxon approximation ( $W$ ) for computing the area under the ROC curve for each classifier for the ATR application. Mean estimates and half-lengths for simultaneous Bonferroni confidence intervals based on 33 independent test data sets (normal assumption with  $\alpha_{total} = 0.05$ ).

Classifier	AUC	$A_z$	$W$
Linear	$0.7866 \pm 0.0013$	$0.6776 \pm 0.0461$	$0.7797 \pm 0.0044$
Quadratic	$0.8833 \pm 0.0070$	$0.8850 \pm 0.0073$	$0.9057 \pm 0.0045$
MLP	$0.9146 \pm 0.0089$	$0.9178 \pm 0.0083$	$0.9130 \pm 0.0100$

(Figure 5.2) reveals that there is something peculiar about the linear ROC curve. Except for two decision thresholds corresponding to the two ROC points, ( $P_{FP} = 0, P_{TP} = 0$ ) and ( $P_{FP} = 1, P_{TP} = 1$ ), the rest of the 101 decision thresholds yield ROC points for the linear classifier concentrated about the Bayes' optimal point ( $P_{FP} = 0.424, P_{TP} = 0.992$ ). Since the linear ROC curve is composed then of essentially only 4 or 5 points, the trapezoidal approximation for AUC may not be a good estimate for the area under the curve. Table 5.6 provides a comparison of AUC and two other area calculation methods, the binormal approximation  $A_z$  (Equation 2.38) and the Wilcoxon approximation  $W$  (Equation 2.39). Table 5.6 shows that the three different methods for calculating the area under the ROC curve for the linear classifier yield three statistically different results. However, area calculations using the trapezoidal method and the binormal approximation are statistically equivalent for the quadratic classifier and all three area computation methods for the MLP classifier are statistically equivalent. These results suggest that the area under the ROC curve may not be a suitable performance measure for the linear classifier and possibly even for the quadratic classifier for this application.

Using average metric distance from the diagonal line as the performance measure, Table 5.5 indicates that the linear classifier is the best classifier. Closer examination of the average ROC curves (Figure 5.2) provides an explanation. As mentioned above, the highest concentration of points that comprise the ROC curve for the linear classifier occur in a very small region compared to the majority of points that comprise the ROC curves for the quadratic and MLP classifiers. This high concentration of points in a small region for the linear classifier reflects the robustness



Table 5.7 Comparison of the classifiers using the modified metric distance.

Classifier	modified metric distance
Linear	$0 \pm 0.0$
Quadratic	$0.5524 \pm 0.0141$
MLP	$0.6057 \pm 0.0106$

of the linear classifier's performance for various decision thresholds. Since the metric distance is the average taxi-cab distance from each ROC point  $(P_{FP}^{(f)}(\theta_i), P_{TP}^{(f)}(\theta_i))$  on ROC curve  $f$  to its corresponding point  $(\theta_i, \theta_i)$  on the diagonal line  $g$ , it is not surprising that the metric distance for the linear statistical classifier is statistically the largest. The choice of the linear classifier using the average metric distance then, represents a choice for consistency in performance.

If  $P_0$  is the maximum acceptable probability of false positive, the proposed metric performance measure (Equation 3.24) can be modified to include a constraint. First define

$$\tilde{\Theta}^f = \left\{ \theta \in \Theta : P_{FP}^{(f)}(\theta) < P_0 \right\} \quad (5.2)$$

Then the modified average metric distance can be defined as

$$\text{modified avg metric distance} = \frac{\int_{\tilde{\Theta}^f} \rho_1(\mathbf{P}^{(f)}(\theta), \mathbf{P}^{(g)}(\theta)) d\theta}{\mu(\Theta)} \quad (5.3)$$

For a finite number  $m$  decision thresholds the modified average metric distance from the ROC curve  $f$  to the diagonal line  $g$  can be approximated as

$$\text{avg metric distance} \approx \frac{\sum_{i=1}^m \rho_1(\mathbf{P}^{(f)}(\theta_i), \theta_i) \text{ such that } P_{FP}^{(f)}(\theta_i) < P_0}{m} \quad (5.4)$$

Table 5.7 shows the comparison of the competing classifiers using this modified performance measure with  $P_0 = 0.3$ . With a maximum acceptable probability of false positive of 0.3, Table 5.7 indicates that the MLP is the best classifier, while the linear classifier is now the worst.

Table 5.8 Estimates for the probability of being the best classifier given the target class for the ATR application. Mean estimates and half-lengths for Bonferroni confidence intervals based on 30 different test data sets.

Classifier	$P_{best}(\text{confuser})$	$P_{best}(\text{target})$
Linear	$0.5422 \pm 0.0000$	$0.9920 \pm 0.0023$
Quadratic	$0.2319 \pm 0.0115$	$0 \pm 0.00$
MLP	$0.2259 \pm 0.0115$	$0.0080 \pm 0.0023$

Table 5.9 Updated estimates and Bonferroni confidence intervals ( $\alpha_{total} = 0.05$ ) for the total and individual class probabilities of the quadratic and MLP classifiers being the best for the ATR application after the linear classifier is removed from consideration.

Classifier	$P_{best}(\text{total})$	$P_{best}(\text{confuser})$	$P_{best}(\text{target})$
Quadratic	$0.6862 \pm 0.0133$	$0.7741 \pm 0.0108$	$0.5596 \pm 0.0287$
MLP	$0.3138 \pm 0.0133$	$0.2259 \pm 0.0108$	$0.4404 \pm 0.0287$

Using the proposed multinomial performance measure  $P_{best}(\text{total})$  as the performance measure, Table 5.5 indicates that the linear classifier is the best classifier. Examination of the estimates for  $P_{best}$  for each class (confuser and target), Table 5.8, provides some insight. Table 5.8 implies that the linear classifier is the most confident (since  $P_{best}$  for each class is the measure on average of which classifier has the highest estimated posterior probability) of the three classifiers for identifying both the confusers and the targets. However, Table 5.8 indicates that linear classifier is the more confident in identifying targets than confusers. Examination of the actual estimated posterior probabilities generated by the linear classifier provides an explanation. The linear classifier has estimated posterior probabilities very close to one for identifying all the targets and confusers, except for one specific confuser. For the serial number E71 truck, the linear classifier has estimated posterior probabilities very close to zero, thereby misclassifying all the data points associated with the truck as a target. If the linear classifier is removed from consideration because this type of performance is unacceptable, then the multinomial selection procedure can be re-accomplished with only the quadratic and MLP classifiers. Table 5.9 shows the updated total probabilities and class probabilities for the quadratic and MLP classifiers being the best. After the linear classifier is removed from consideration, Table 5.9 implies that the quadratic classifier is the best. Closer examination of Table 5.9 indicates that the quadratic classifier is more confident than the MLP

classifier for identifying both the confusers and the targets, but especially more confident for identifying confusers. The choice of the quadratic classifier over the MLP classifier in this two classifier comparison represents a choice for the classifier with the greater confidence, i.e., statistically larger estimated class posterior probabilities.

## 5.6 Conclusions

If a high false alarm rate as high as 0.42 is acceptable, then the linear classifier is considered the best classifier according to both the proposed metric distance and the multinomial selection procedure. If  $P_{FP} = 0.42$  is not acceptable, then the modified metric distance would provide support for the MLP as the best classifier. The choice of the best classifier then depends upon a thorough understanding of the performance measures and the desired requirements for the specific application.

## *VI. Pilot Workload Application*

### *6.1 Overview*

This chapter provides comparisons between the methodologies introduced in Chapter III and IV and typical approaches on a pilot workload application. As discussed in Section 1.2.2, the USAF is especially interested in pilot workload detection because pilot overload or task saturation can decrease mission effectiveness and, in some extreme cases, cause loss of life [9]. In order for pilots to be confident in classification systems that may be employed in their cockpits in the future, they must have an objective way of testing and evaluating competing systems. The Air Force Research Laboratory Human Effectiveness (AFRL/HE) Division at Wright-Patterson AFB is leading the effort toward modeling pilot workload using psychophysiological measures and EEG. The application summarized in this chapter uses two statistical classifiers and an ANN classifier on data provided by AFRL/HE to illustrate the methodologies developed in this dissertation.

The research in this chapter has its foundation in a technical report [1] and a journal article [14]. This chapter is organized as follows. Section 6.2 provides a description of the pilot workload data used. Section 6.3 describes the experimental setup and classifiers employed. Section 6.4 provides the results obtained using the various methodologies to compare the competing classifiers. Discussion of the results are given in Section 6.5 and conclusions are provided in Section 6.6.

### *6.2 Data Description*

The data set used in this application is taken from a pilot workload study conducted by AFRL/HE using ten different pilots of the Wright-Patterson AFB Aero Club as test subjects. Each pilot flew two missions on two different days following the same flight profile. The profile consisted of pre-flight preparations and a take-off from the home field, VFR (visual flight rule) departure, IFR (instrument flight rule) arrival, descent, and instrument approach at the alternate field followed by a landing and subsequent return to the home field containing both IFR and VFR

flight segments. This profile is divided into 22 two minute segments which are each designated by instructors prior to the study as either low, medium, or high workload. For example, the pre-flight is designated as low workload, while the IFR mid-flight cruise is designated as medium workload, and the landings as high workload.

Throughout the flight, EEG are sampled at a frequency of 256 Hz at 29 different locations on the pilot's head. Bad data associated with *artifacts* such as head movements and muscle movement which can be caused by speech are replaced with the average of data that is collected before and after the *artifact*. The EEG data are passed through a fast-Fourier transform (FFT) and then broken up into five bands based upon frequency:  $\Delta$ ,  $\theta$ ,  $\alpha$ ,  $\beta$ , and  $\mu\beta$ . For each band at each head location, the average power is collected over a 10-second moving window, resulting in 23 data points consisting of 145 EEG features ( $5 \text{ bands} \times 29 \text{ locations}$ ) for each two minute segment of flight [26].

Six peripheral psychophysiological features are also collected during the flight. Cardiopulmonary information is represented by heart rate (HR) measured in number of beats per minute and heart interbeat interval (IBI) which is a measure of the variability of the heart rhythm. Eye information is represented by eye blinks (BLK) measured in number of eye blinks per millisecond and interblink interval (IBLKI) which is the time between blinks. Respiratory information is represented by the breathing rate (BTH) measured in number of breaths per millisecond and interbreath interval (IBTHI) which is the time between breaths. For each peripheral feature a 10-second moving window is also used to extract 23 data points for each two minute segment of flight [26].

The multivariate data set contains 506 data points (23 points for each 2-minute flight segment) consisting of 151 features (145 EEG and 6 peripheral features) collected for each pilot for each day. For this application, only the data for one pilot on one day is used to illustrate the methodologies developed in this dissertation.

### 6.3 Experiment Description

**6.3.1 Experimental setup.** Since the USAF is especially interested in identifying pilot overload conditions, this experiment classifies pilot workload into two classes: non-overload or low and medium workload (class 1) and overload or high workload (class 2). This division of classes results in an unbalanced data set with 299 data points (59.1% of the data) belonging to the non-overload class, while only 207 points (40.9%) belong to the overload class. This data set has a large number of features and a limited number of data points. According to Foley's rule of thumb [31], the test set error is close to the optimum training error attained by a Bayes classifier for inputs with unknown distributions when

$$M_{train} \gg 3 \cdot I \cdot K \quad (6.1)$$

where  $M_{train}$  is the number of training data points,  $I$  is the number of features, and  $K$  is the number of classes. Since the number of total points in this data set (506) violates Foley's rule, the number of features need to be reduced. The SNR screening method (Figure 6.1) is used to identify the salient features shown in Table 6.1.

All classifiers are trained on approximately 50% of the data and tested on the remaining 50% using these salient features. Both training and testing data sets maintain the same proportion of class 1 and class 2 data. In order to ensure that the performance of the classifiers are not dependent upon a particular choice of training and testing data sets, the data are *shuffled* 30 times to generate 30 random selections of the training and testing data sets.

**6.3.2 Classifiers.** A linear statistical classifier, a quadratic statistical classifier, and an ANN are used. The linear statistical classifier is employed using a discriminant analysis function using equal (pooled) covariance matrices for each class while the quadratic statistical classifier employs unequal covariance matrices for each class. The neural network employed is a MLP

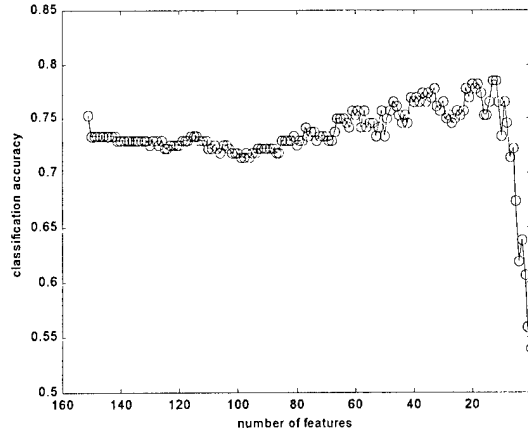


Figure 6.1 Plot of classification accuracy (on independent test data set) vs. number of features generated by the Signal-to-Noise Ratio (SNR) algorithm [14] for the pilot workload application.

Table 6.1 Listing of 14 most salient features identified by SNR algorithm for pilot workload application.

SNR Saliency Ranking	Feature Number	Description of Feature
1	19	average power of $\beta$ band at location $C6$
2	119	average power of $\beta$ band at location $P9$
3	129	average power of $\beta$ band at location $P04$
4	23	average power of $\alpha$ band at location $CZ$
5	146	number of heart beats per minute
6	145	average power of $\mu\beta$ band at location $T8$
7	16	average power of $\Delta$ band at location $C6$
8	128	average power of $\alpha$ band at location $P04$
9	132	average power of $\theta$ band at location $PZ$
10	18	average power of $\alpha$ band at location $C6$
11	111	average power of $\Delta$ band at location $P8$
12	150	number of breaths per millisecond
13	20	average power of $\mu\beta$ band at location $C6$
14	105	average power of $\mu\beta$ band at location $P4$

trained using MATLAB's adaptive learning algorithm (TRAINGDX) with an initial learning rate of 0.01 [23]. This algorithm also employs momentum with a momentum constant of 0.9. One hidden layer is employed with 38 nodes. All activation functions are sigmoidal. Thirty statistical classifiers for each type and thirty neural networks are trained using 254 randomly selected data points (59.1% class 1/ 40.9% class 2) for training. The neural networks use 40% of this training data for internal validation to prevent over training. All the classifiers use 252 randomly selected data points (59.1% class 1/ 40.9% class 2) for independent testing of their performance. All features are standardized to zero mean and unit variance. Each neural network uses a different random initialization for the weights.

#### 6.4 Results

The confusion matrices for all three classifiers are shown in Tables 6.2-6.4. These confusion matrices represent classifier performance at Bayes optimal point, i.e., classification performance at a decision threshold that minimizes total classification error. The raw numbers in the confusion matrix are the sums over all 30 test data sets. The percentages are the means of 30 probabilities conditioned on the rows for each test data set. Bonferroni confidence intervals are computed using the normal assumption with  $\alpha_{total} = 0.05$ .

ROC curves for each classifier are generated by using 101 discrete decision thresholds for each test data set. The average (over 30 test data sets) ROC curves are shown in Figure 6.2. The area under the ROC curve (AUC) for the three classifiers is estimated by using the trapezoidal method for all 30 ROC curves for each classifier and computing the mean. The average (over all thresholds) of the proposed ROC metric  $d_{\rho_1,1}$  (Section 3.4) is computed for all 30 ROC curves to compare each classifier to the  $\theta = P_{TP} = P_{FP}$  diagonal (chance) line for each test data set. The mean metric distances along with the mean AUCs are shown in Table 6.5.



Table 6.2 Linear classifier confusion matrix for Pilot Workload application. Raw numbers are the sums over 30 test data sets. Percentages and corresponding half-lengths of the confidence intervals (CIs) are also based on 30 different test data sets (Bonferroni CIs using normal assumption with  $\alpha_{total} = 0.05$ ).

Linear Classifier CLASSIFICATION			
T R U T H		non-Overload	Overload
	non-Overload	3170 (70.92 $\pm$ 1.98%)	1300 (29.08 $\pm$ 1.98%)
	Overload	666 (21.55 $\pm$ 1.33%)	2424 (78.45 $\pm$ 1.16%)

Table 6.3 Quadratic classifier confusion matrix for Pilot Workload application. Raw numbers are the sums over 30 test data sets. Percentages and corresponding half-lengths of the confidence intervals are also based on 30 different test data sets (Bonferroni CIs using normal assumption with  $\alpha_{total} = 0.05$ ).

Quadratic Classifier CLASSIFICATION			
T R U T H		non-Overload	Overload
	non-Overload	3806 (85.15 $\pm$ 1.06%)	664 (14.85 $\pm$ 1.06%)
	Overload	945 (30.58 $\pm$ 2.21%)	2145 (69.42 $\pm$ 2.21%)

Table 6.4 MLP classifier confusion matrix for Pilot Workload application. Raw numbers are the sums over 30 test data sets. Percentages and corresponding half-lengths of the confidence intervals are also based on 30 different test data sets (Bonferroni CIs using normal assumption with  $\alpha_{total} = 0.05$ ).

MLP Classifier CLASSIFICATION			
T R U T H		non-Overload	Overload
	non-Overload	3765 (84.23 $\pm$ 1.61%)	705 (15.77 $\pm$ 1.61%)
	Overload	654 (21.17 $\pm$ 1.91%)	2436 (78.83 $\pm$ 1.91%)

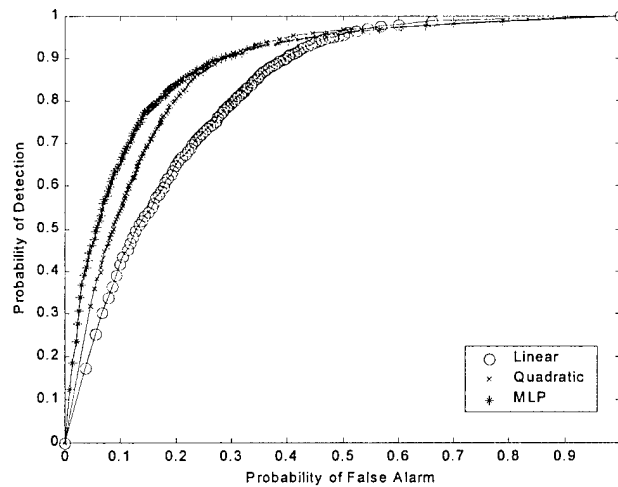


Figure 6.2 Average ROC curves for Pilot Workload application. Curves averaged over 30 different test data sets.

The BEM multinomial selection procedure as described in Section 4.2 is also applied to the pilot workload problem. The conditional probabilities for each classifier being the best given the class (overload or non-overload) are computed for each test data set. The prior probability of overload (or non-overload) is then estimated for each test data set by computing the ratio of the number of overload (or non-overload) workload data points to the total number of pilot workload points. These prior probability estimates are approximately the same for each test data set since the same proportion of class 1 and class 2 data (59.1% class 1/ 40.9% class 2) are maintained for all training and testing data sets. Using these estimates for the prior probabilities, the total probability that each classifier is the best for each test data set is computed using the law of total probability (Equation 4.2). The mean (over 30 test data sets) total probabilities are also given in Table 6.5.

Table 6.5 also contains estimates for classification accuracy which is the typical performance index used in pattern recognition. Classification accuracy (CA) is estimated for a specific decision threshold, namely the Bayes optimal point. The Bayes optimal point is the decision threshold for which the total misclassification error is a minimum. Classification accuracy is defined for this

Table 6.5 Summary of various performance measures for Pilot Workload application. Mean estimates and half-lengths for Bonferroni confidence intervals (normal assumption with  $\alpha_{total} = 0.05$ ) based on 30 different test data sets.

Classifier	CA	AUC	Metric Distance	$P_{best}(total)$
Linear	$0.7399 \pm 0.0103$	$0.8247 \pm 0.0100$	$0.4513 \pm 0.0109$	$0.1877 \pm 0.0149$
Quadratic	$0.7872 \pm 0.0100$	$0.8687 \pm 0.0088$	$0.5187 \pm 0.0161$	$0.5716 \pm 0.0153$
MLP	$0.8202 \pm 0.0125$	$0.8884 \pm 0.0114$	$0.5214 \pm 0.0276$	$0.2407 \pm 0.0278$

application as follows:

$$CA = \frac{\text{number of overload and non-overload workload points classified correctly}}{\text{total number of pilot workload data points}} \quad (6.2)$$

Classification accuracy is computed for each test data set for each classifier. The mean (over 30 test data sets) CAs are reported in Table 6.5.

For all the performance measures shown in Table 6.5, the half-lengths for confidence intervals are also reported. These half-lengths represent 95% simultaneous Bonferroni [46] confidence intervals for comparing the three different classifiers.

### 6.5 Discussion

Using classification accuracy (CA) as the performance measure, Table 6.5 indicates that the MLP is the best classifier. However, when AUC and the average proposed metric distance are used as the performance measures, the best classifier is unclear from the table. The confidence intervals for the quadratic statistical classifier and the MLP overlap for both AUC and metric distance. An examination of the average ROC curves (Figure 6.2) provides an explanation. The ROC curve for the MLP dominates the ROC curve for the quadratic statistical classifier for  $P_{FP}$  (probability of false positive/false alarm) values from 0 to 0.3, while the quadratic ROC curve dominates the MLP ROC curve for  $P_{FP}$  values from 0.3 to 0.6. For  $P_{FP}$  values from 0.6 to 1.0, the two ROC curves

Table 6.6 Estimates for the probability of being the best classifier given the workload class for Pilot Workload application. Mean estimates and half-lengths for Bonferroni confidence intervals based on 30 different test data sets.

Classifier	$P_{best}(non - overload)$	$P_{best}(overload)$
Linear	$0.0899 \pm 0.0126$	$0.3291 \pm 0.0411$
Quadratic	$0.7027 \pm 0.0265$	$0.3819 \pm 0.0210$
MLP	$0.2074 \pm 0.0252$	$0.2890 \pm 0.0436$

coincide. Therefore, it is not surprising that the AUCs are statistically equivalent for the quadratic statistical classifier and the MLP. A further examination of the ROC curves, indicates that the highest concentration of points that comprise both the MLP and the quadratic ROC curves occurs in the same region ( $P_{FP}$  values from 0.1 to 0.3). Since the metric distance is the average taxi-cab distance from each ROC point, it is also not surprising that the metric distances for the MLP and the quadratic statistical classifier are statistically equivalent.

The question that naturally arises after examining Table 6.5 is, *Why does the proposed multinomial performance measure  $P_{best}(total)$  indicate that the quadratic statistical classifier is the best?* Examination of the estimates for  $P_{best}$  for each workload class (overload and non-overload), Table 6.6, provides some insight. Table 6.6 implies that the quadratic classifier is the most confident (since  $P_{best}$  for each class is the measure on average of which classifier has the highest estimated posterior probability) of the three classifiers for identifying the non-overload workload class. Since the confidence intervals for  $P_{best}(overload)$  estimates for the three classifiers overlap, it is not clear which classifier is the most confident of the three for identifying the overload class. However, since  $P_{best}(total)$  is the weighted average of  $P_{best}$  for each workload class (overload and non-overload), it is not surprising that  $P_{best}(total)$  indicates that the quadratic statistical classifier is the best overall for identifying both workload classes.

## 6.6 Conclusions

Since the quadratic and MLP classifiers have statistically equivalent performance measures for AUC and metric distance, this application demonstrates the usefulness of an alternate method

like the multinomial selection procedure for differentiating between the competing classifiers. The multinomial selection procedure could be used as a tie-breaker for determining the best classifier when the other methods cannot select a winner.

## VII. An Interpretation of Performance Measures

### 7.1 Overview

This chapter provides an interpretation of the performance measures used in the two previous application chapters. Sections 7.2 and 7.3 review the interpretation of classification accuracy and area under the ROC curve which are the typical performance measures used to compare competing classifiers. Sections 7.4 and 7.5 explore the interpretation of the average metric distance from the diagonal and the probability of being the best as suggested by their use in the two applications. Finally, Section 7.6 provides a summary of the interpretations of the new performance measures introduced in this dissertation along with the interpretations of the typical performance measures.

### 7.2 Interpretation of Classification Accuracy

Consider the confusion matrix shown in Table 7.1 for a two-class problem. The number of Class 1 exemplars classified correctly is  $N_{1C}$  and the number of Class 2 exemplars classified correctly is  $N_{2C}$ . Alternatively, the number of Class 1 exemplars classified incorrectly is  $N_{1\bar{C}}$  and the number of Class 2 exemplars classified incorrectly is  $N_{2\bar{C}}$ . For  $n_1$  Class 1 exemplars and  $n_2$  Class 2 exemplars, the estimated probability of successful classification or classification accuracy  $CA$  is given by

$$CA = \frac{N_{1C} + N_{2C}}{n_1 + n_2}. \quad (7.1)$$

Classification accuracy indicates how well the classifier is at identifying both Class 1 and Class 2 exemplars at a specific decision threshold  $\theta$ . Typically, classification accuracy is reported for

Table 7.1 Example Confusion Matrix for computing classification accuracy.

Actual Membership	Assigned Membership		
	Class 1	Class 2	totals
	Class 1	Class 2	totals
	$N_{1C}$	$N_{1\bar{C}}$	$n_1$
	$N_{2\bar{C}}$	$N_{2C}$	$n_2$

the Bayes optimal point. The Bayes optimal point is the decision threshold for which the total misclassification error ( $1 - CA$ ) is a minimum.

Let  $z \in \Re$  be a scalar that represents some measure of the classifier's strength of conviction for Class 1, such that  $z > \theta$  for an exemplar generates an assignment of Class 1 while  $z < \theta$  for an exemplar generates an assignment of Class 2. Let  $Z_1$  and  $Z_2$  be random variables representing the values of  $z$  for a randomly selected exemplar from Class 1 and Class 2, respectively. For the case where a classifier with a decision threshold  $\theta$  is presented with one randomly chosen Class 1 exemplar and one randomly chosen Class 2 exemplar, the probability that the classifier will successfully identify both exemplars is given by [20]:

$$P(S) = \Pr(Z_1 > \theta) \Pr(Z_2 < \theta) \quad (7.2)$$

Classification accuracy is effectively an estimate of  $P(S)$ .

### 7.3 Interpretation of Area Under the ROC Curve

The area under the ROC curve (AUC) represents the probability that a randomly chosen target exemplar is correctly rated with greater suspicion than a randomly chosen non-target exemplar [20]. If  $z$ , as defined in the previous section, is used as the rating and Class 2 exemplars are the targets while Class 1 exemplars are the non-targets, then lower values of  $z$  equate to stronger indications of target. The probability  $P(CR)$  that a randomly chosen target exemplar is correctly rated with greater suspicion than a randomly chosen non-target exemplar is then given by [20]

$$P(CR) = \Pr(z_2 < z_1). \quad (7.3)$$

This probability of correct rating  $P(CR)$  is the same quantity estimated by the Wilcoxon statistic  $W$  (Section 2.3.3, page 2.39), which is sometimes used to estimate AUC. Therefore, AUC effectively measures  $P(CR)$  which is independent of the decision threshold  $\theta$ .

#### 7.4 Interpretation of Average Metric Distance from Diagonal

The applications in the previous chapters suggest that the average metric distance between the ROC curve and the diagonal line reflects the robustness of the classifier's performance for various decision thresholds. This robustness hypothesis is tested in this section on a variety of data sets. Since it is suspected that the perturbation of data points of one class is equivalent to a small change in decision threshold, classifier performance as a function of perturbation of Class 2 data is examined. Specifically, classification accuracy ( $CA$ ) for each classifier is computed for various levels of perturbation of Class 2 data and is compared to  $CA$  changes expected based on the values of the average metric distance from the diagonal line. Classifiers with similar values for the average metric distance are expected to have similar changes in  $CA$  as a function of perturbation level. Alternatively, a classifier with a higher value for the average metric distance from the diagonal is expected to have  $CA$  more stable with changes in perturbation level than classifiers with smaller values for the average metric distance.

**7.4.1 2-D Normal Data Set.** Consider the following 2-D normal data set consisting of 1000 randomly generated Class 1 data points and 1000 randomly generated Class 2 data points. Class 1 data are generated from a two-dimensional normal distribution with mean  $\vec{\mu}_1$  and covariance matrix  $\Sigma_1$  given by

$$\vec{\mu}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{and} \quad \Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad (7.4)$$



Table 7.2 Classification accuracy and average metric distance from diagonal for unperturbed 2-D normal data.

Classifier	CA	Metric Distance
Linear	0.770	0.362
Quadratic	0.766	0.350
MLP	0.768	0.347

while Class 2 data are generated from a two-dimensional normal distribution with mean  $\vec{\mu}_2$  and covariance matrix  $\Sigma_2$  given by

$$\vec{\mu}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \text{and} \quad \Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \quad (7.5)$$

The data are classified using three different types of classifiers: a linear statistical classifier, a quadratic statistical classifier, and a multi-layer perceptron (MLP) artificial neural network. All three classifiers are trained on fifty percent of the data (balanced between the two classes) and tested and compared on the remaining fifty percent. The linear statistical classifier is employed using a discriminant analysis function using equal (pooled) covariance matrices for each class while the quadratic statistical classifier employs unequal covariance matrices for each class. The MLP is trained using MATLAB's adaptive learning algorithm (TRAINGDX) with an initial learning rate of 0.01 [23]. One hidden layer is employed with eight nodes. Forty percent of the training data is used for internal validation of the MLP to prevent over training. The classification accuracy is computed using Equation 7.1 for all three classifiers as shown in Table 7.2. After generating ROC curves (Figure 7.1) for each classifier using 101 discrete decision thresholds, the average metric distance from the diagonal line is computed for all three classifiers and is also shown in Table 7.2.

Since the average metric distances from the diagonal in Table 7.2 are all within 0.01 of each other, all three classifiers are expected to have similar variations in  $CA$  when tested on perturbed data. In order to test this hypothesis, the trained classifiers are tested on new test data sets formed by perturbing the Class 2 data points in the original test data set. Class 2 data points  $\vec{x}_i$

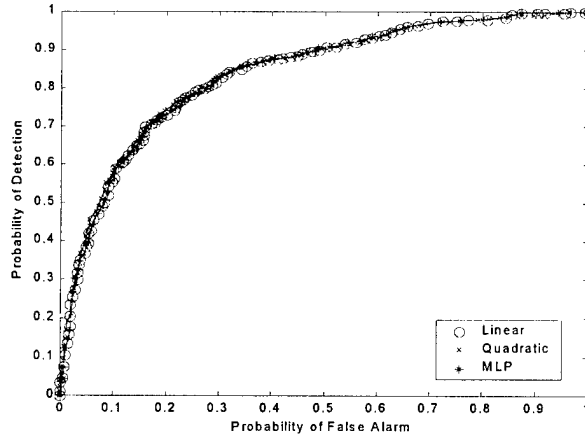


Figure 7.1 ROC curves for unperturbed 2-D Normal data.

are perturbed by moving the points in the direction defined by the vector connecting the means of Class 1 and Class 2 as follows:

$$\vec{x}_i^{perturbed} = \vec{x}_i + \lambda \cdot (\vec{\mu}_2 - \vec{\mu}_1) \quad (7.6)$$

where  $\lambda$  is a fraction indicating the amount of perturbation. All three trained classifiers are tested on the perturbed data sets formed by varying the perturbation fraction  $\lambda$  from  $-0.6$  to  $+0.6$ . Classification accuracies for the classifiers are plotted as a function of perturbation fraction  $\lambda$  in Figure 7.2. As expected all three classifiers have similar variations in  $CA$  on the perturbed data.

*7.4.2 University of Wisconsin Breast Cancer Diagnosis Data Set.* Consider the University of Wisconsin Breast Cancer Diagnosis Data Set as used in Experiment #2 in Section 3.4.3. Recall that this data set consists of 699 patterns of which 458 are benign samples and 241 are malignant samples. Each of these patterns consists of nine measurements taken from fine needle aspirates from a patient's breast. However, for this experiment only the top two ranked features (bare nuclei and clump thickness) are used to construct a multi-layer perceptron (MLP) artificial neural network, a linear statistical classifier, and a quadratic statistical classifier. Thirty classifiers of each type

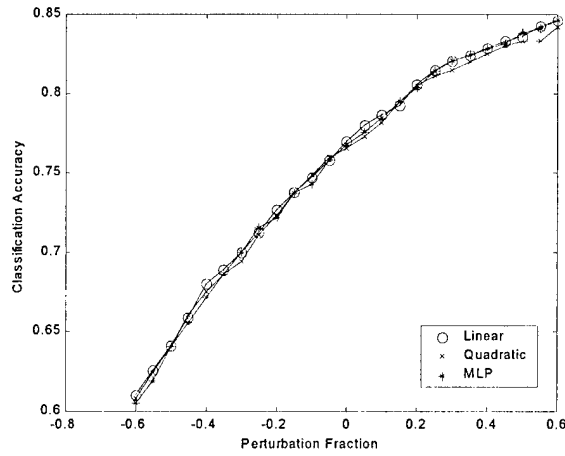


Figure 7.2 Classification accuracy as a function of perturbation fraction for linear, quadratic, and MLP classifiers for 2-D normal data set.

Table 7.3 Mean classification accuracies and average metric distances from diagonal line for linear, quadratic, and MLP classifiers. Bonferroni confidence intervals based on 30 different test data sets (normal assumption with  $\alpha_{total} = 0.05$ ).

Classifier	CA	Metric Distance
Linear	$0.935 \pm 0.004$	$0.763 \pm 0.005$
Quadratic	$0.941 \pm 0.004$	$0.840 \pm 0.007$
MLP	$0.945 \pm 0.004$	$0.800 \pm 0.010$

are generated and tested by randomly shuffling the data into thirty different training and testing data sets. The mean classification accuracies over the thirty different test data sets for the three classifiers are shown in Table 7.3. ROC curves for each classifier are generated using 101 discrete decision thresholds for each test data set. The average ROC curves over 30 different test data sets for the three different types of classifiers are shown in Figure 7.3. The means over the 30 different test data sets for the average metric distance from the diagonal line is computed for all three classifiers and is also shown in Table 7.3.

Table 7.3 shows that the average metric distances from the diagonal for the quadratic classifier is statistically larger than the average metric distance for the MLP classifier and even larger statistically than the linear classifier. The implication is that the quadratic classifier is expected to

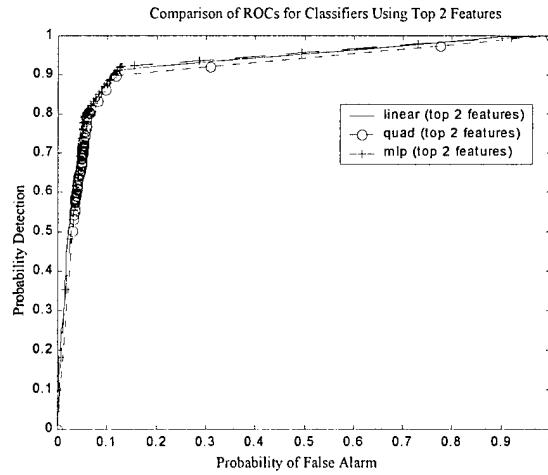


Figure 7.3 Average ROC curves for University of Wisconsin Breast Cancer Diagnosis Data Set. ROC curves averaged over 30 different test data sets.

have less variation in  $CA$  than the other two classifiers when tested on perturbed data. In order to test this hypothesis, the thirty trained classifiers of each type are tested on new test data sets formed by perturbing the Class 2 data points in the original test data set associated with each trained classifier. The Class 2 data points for each test data set are perturbed as described in Section 7.4.1 by varying  $\lambda$  from  $-0.8$  to  $+0.6$ . The mean classification accuracies over the thirty different perturbed test data sets for the classifiers are plotted as a function of the perturbation fraction  $\lambda$  in Figure 7.4. As expected the quadratic classifier displays the least variation in  $CA$  on the perturbed data.

**7.4.3 ATR Data Set.** Consider the ATR data set described in Sections 5.2-5.3. Recall that the ATR application classifies targets into two classes: non-targets or confusers (class 1) and targets specified for attack (class 2). Furthermore, a linear statistical classifier, a quadratic statistical classifier, and a multi-layer perceptron (MLP) artificial neural network are trained on one set of data and then tested on thirty-three independent test data sets. For one particular test data set, the classification accuracies for the three classifiers are shown in Table 7.4. After generating ROC curves (Figure 7.5) for each classifier on this particular test data set using 101

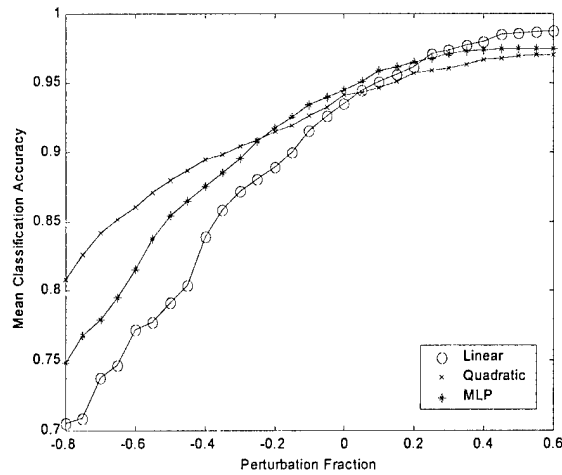


Figure 7.4 Mean classification accuracy as a function of perturbation fraction for linear, quadratic, and MLP classifiers for University of Wisconsin Breast Cancer Diagnosis Data Set.

Table 7.4 Classification accuracy and average metric distance from diagonal for one ATR test data set.

Classifier	CA	Metric Distance
Linear	0.733	0.731
Quadratic	0.839	0.622
MLP	0.849	0.584

discrete decision thresholds, the average metric distance from the diagonal line is computed for all three classifiers and is also shown in Table 7.4.

Table 7.4 shows that the average metric distances from the diagonal for the linear classifier is larger than the average metric distance for the quadratic and MLP classifiers by approximately 0.11 or more. The implication is that the linear classifier is expected to have much less variation in *CA* than the other two classifiers when tested on perturbed data. In order to test this hypothesis, the trained classifiers of each type are tested on new test data sets formed by perturbing the Class 2 data points in the original test data. The Class 2 data points for each test data set are perturbed as described in Section 7.4.1 by varying  $\lambda$  from  $-0.6$  to  $+0.6$ . Classification accuracies for the

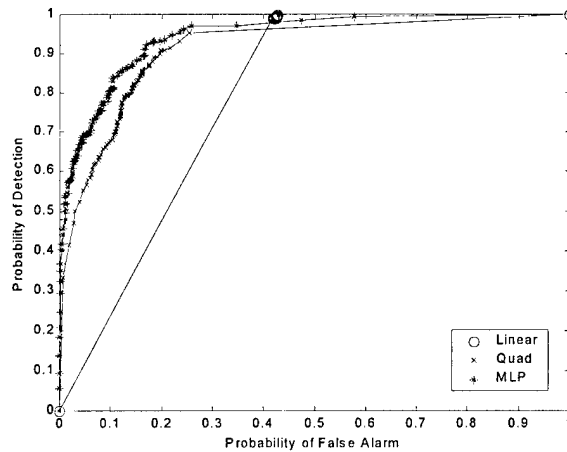


Figure 7.5 ROC curves for one particular ATR test data set.

classifiers are plotted as a function of perturbation fraction  $\lambda$  in Figure 7.6. As expected the linear classifier displays very stable  $CA$  on the perturbed data compared to the other two classifiers.

**7.4.4 Pilot Workload Data Set.** Consider the pilot workload data set described in Sections 6.2-6.3. Recall that the multivariate data set used contains 506 data points consisting of 14 salient features. Furthermore, a linear statistical classifier, a quadratic statistical classifier, and a multi-layer perceptron (MLP) artificial neural network are trained to classify the data into two workload classes: non-overload (class 1) and overload (class 2). Thirty classifiers of each type are generated and tested by randomly shuffling the data into thirty different training and testing data sets. For one particular test data set, the classification accuracies for the three classifiers are shown in Table 7.5. After generating ROC curves (Figure 7.7) for each classifier on this particular test data set using 101 discrete decision thresholds, the average metric distance from the diagonal line is computed for all three classifiers and is also shown in Table 7.5.

Since the average metric distances from the diagonal in Table 7.5 are all within 0.07 of each other, all three classifiers are expected to have similar variations in  $CA$  when tested on perturbed data. In order to test this hypothesis, the trained classifiers of each type are tested on new test

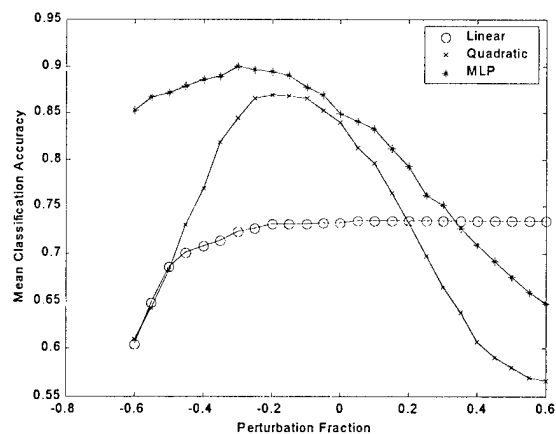


Figure 7.6 Classification accuracy as a function of perturbation fraction for linear, quadratic, and MLP classifiers for one particular ATR test data set.

Table 7.5 Classification accuracy and average metric distance from diagonal for one pilot workload test data set.

Classifier	CA	Metric Distance
Linear	0.762	0.501
Quadratic	0.782	0.507
MLP	0.841	0.571

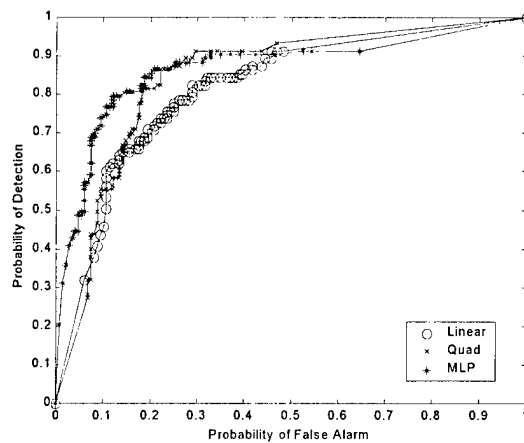


Figure 7.7 ROC curves for one particular pilot workload test data set.

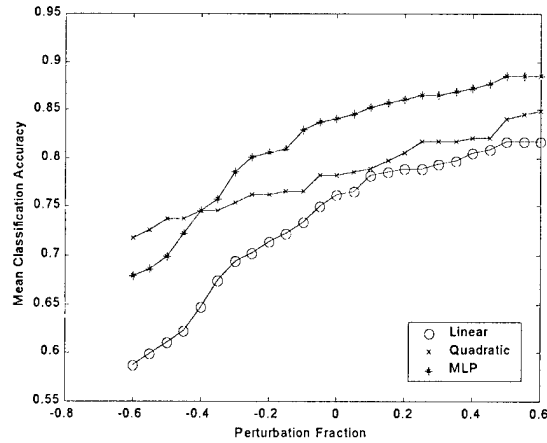


Figure 7.8 Classification accuracy as a function of perturbation fraction for linear, quadratic, and MLP classifiers for one particular pilot workload test data set.

data sets formed by perturbing the Class 2 data points in the original test data. The Class 2 data points for each test data set are perturbed as described in Section 7.4.1 by varying  $\lambda$  from  $-0.6$  to  $+0.6$ . Classification accuracies for the classifiers are plotted as a function of perturbation fraction  $\lambda$  in Figure 7.8. The quadratic classifier appears to have the slightly less variation in  $CA$  on the perturbed data than the linear or MLP classifiers. However, if more test data sets are used and averaged, the curves in Figure 7.8 are expected to smooth out and indicate that all three classifiers have similar variations in  $CA$  on the perturbed data.

### 7.5 Interpretation of Probability of Being the Best

The applications in the previous chapters suggest that the probability of being the best reflects the confidence or strength of conviction of a classifier for its classification of the data. This strength of conviction hypothesis is tested in this section on a couple of 2-D data sets. For 2-D data sets, contour plots of the estimated posterior probabilities for either class can be generated for each classifier for the 2-D feature space. If Classifier A has a larger probability of being the best for a given class than Classifier B, then exemplars from the appropriate class in the 2-D feature space will



Table 7.6 Estimates and Bonferroni confidence intervals ( $\alpha_{total} = 0.05$ ) for the probability of being the best classifier given Class 2 data for 2-D Normal Data Set.

	Linear	Quadratic	MLP
$P_{best}(\text{Class 2})$	0.50	0.44	0.06
CI	[0.45 0.55]	[0.39 0.49]	[0.03 0.09]

have on average higher estimated posterior probabilities for Classifier A than Classifier B. Since the estimated posterior probabilities for a two-class problem are directly related, contour plots for just the target class (Class 2) are used in the following examples to test the strength of conviction hypothesis.

*7.5.1 2-D Normal Data Set.* Consider the same 2-D normal data set described in Section 7.4.1 above. The point estimates for the probability of being the best classifier given Class 2 data, along with their corresponding Bonferroni confidence intervals (using Equation 2.14 with  $\alpha_{total} = 0.05$ ) for comparing the three classifiers, are given in Table 7.6. Since  $P_{best}$  for the linear and quadratic classifiers are greater statistically than the MLP classifier, the expectation is that the linear and quadratic classifiers assign higher estimated Class 2 posterior probabilities for appropriate points in the 2-D feature space than the MLP classifier.

The Class 2 probability contour plots for the three classifiers are shown in Figures 7.9-7.11. A comparison of the three plots indicates that the Class 2 data points have on average higher estimated posterior probabilities according to the contour lines of the linear and quadratic classifiers than the contour lines of the MLP classifier.

*7.5.2 University of Wisconsin Breast Cancer Diagnosis Data Set.* Consider the same University of Wisconsin Breast Cancer Diagnosis Data Set described in Section 7.4.2 above. The means over the thirty different test data sets for the probability of being the best classifier given Class 2 data, along with their corresponding Bonferroni confidence intervals (Normal assumption with  $\alpha_{total} = 0.05$ ) for comparing the three classifiers, are given in Table 7.7. Since  $P_{best}$  for the quadratic classifier is much greater statistically than both the linear and MLP classifiers, the

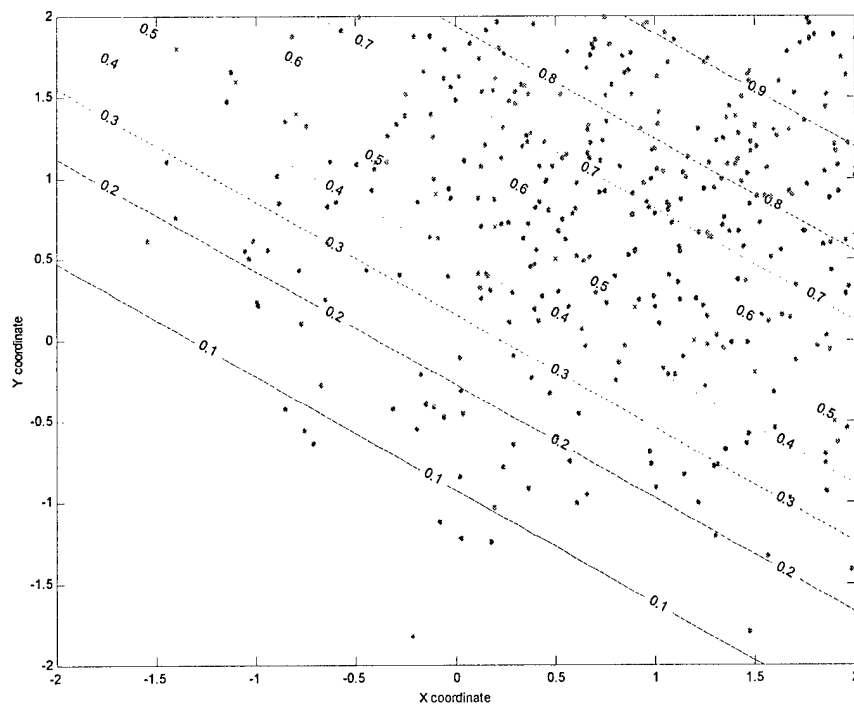


Figure 7.9 Estimated Linear posterior probability contour plot for 2-D Normal Data Set. (Dots represent actual Class 2 data points and X's mark class boundary).

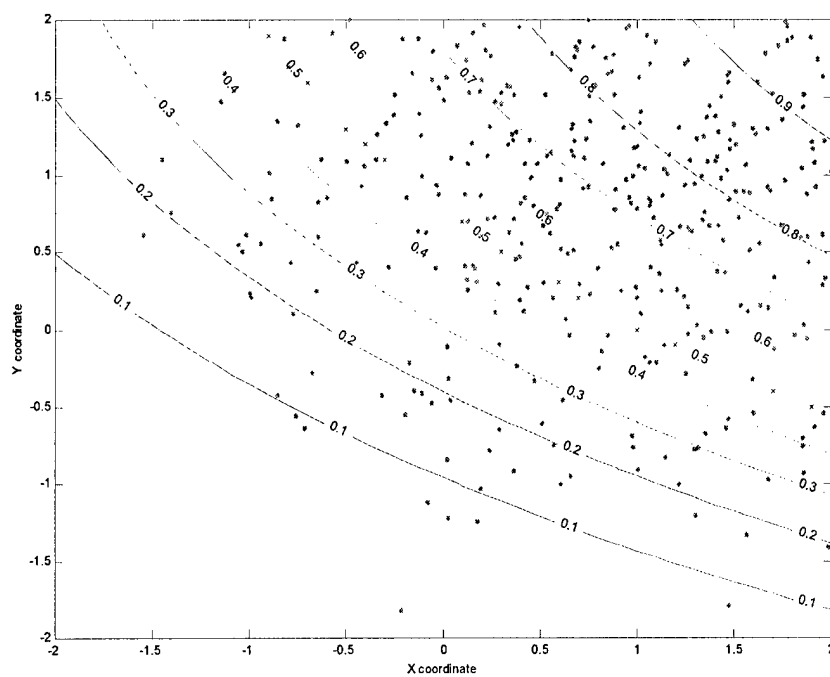


Figure 7.10 Estimated Quadratic posterior probability contour plot for 2-D Normal Data Set. (Dots represent actual Class 2 data points and X's mark class boundary).

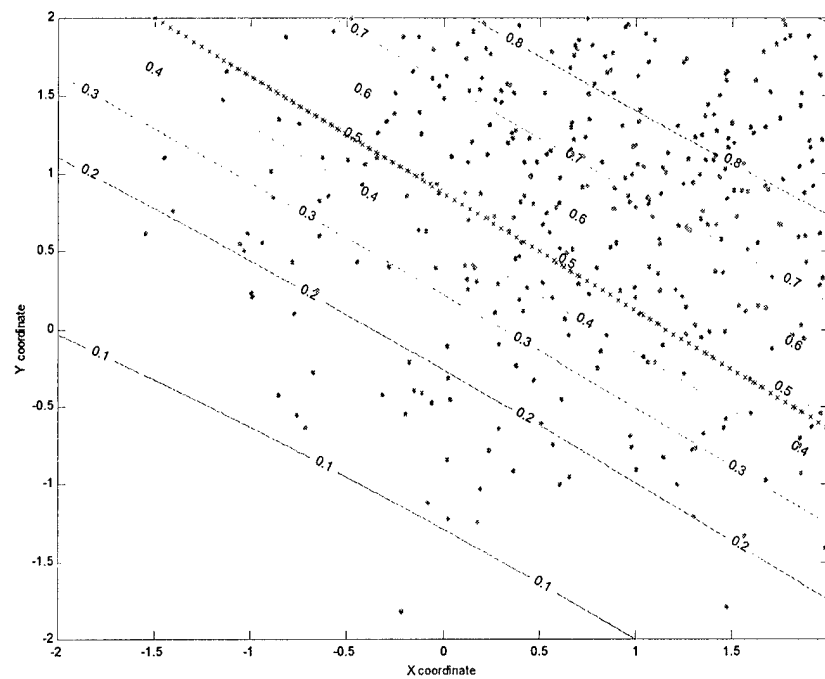


Figure 7.11 Estimated MLP posterior probability contour plot for 2-D Normal Data Set. (Dots represent actual Class 2 data points and X's mark class boundary).

Table 7.7 Means and Bonferroni confidence intervals ( $\alpha_{total} = 0.05$ ) for the probability of being the best classifier given Class 2 data for University of Wisconsin Breast Cancer Diagnosis Data Set.

	Linear	Quadratic	MLP
$P_{best}(\text{Class 2})$	0.01	0.87	0.12
CI	[0.00 0.02]	[0.85 0.89]	[0.10 0.14]

expectation is that the quadratic classifier assigns higher estimated Class 2 posterior probabilities for appropriate points in the 2-D feature space than the other two classifiers.

The Class 2 probability contour plots for the three classifiers are shown in Figures 7.12-7.14. A comparison of the three plots indicates that the Class 2 data points have on average higher estimated posterior probabilities according to the contour lines of the quadratic classifier than the contour lines of the of the linear and MLP classifiers.

## 7.6 Conclusions

This chapter reviews the interpretation of the typical performance measures used to compare competing classifiers and explores the interpretation of the new performance measures introduced in this dissertation. Classification accuracy indicates how well the classifier is at identifying exemplars from all classes at a specific decision threshold. For a two-class problem, where a classifier, operating at a specific decision threshold, is presented with one randomly chosen Class 1 exemplar and one randomly chosen Class 2 exemplar,  $CA$  is effectively the probability that the classifier will successfully identify both exemplars. The area under the ROC curve is independent of decision threshold and represents the probability that a randomly chosen target (Class 2) exemplar is correctly rated with greater suspicion than a randomly chosen non-target (Class 1) exemplar. The examples in this chapter support the hypothesis that the average metric distance between the ROC curve and the diagonal line reflects the robustness of the classifier's performance for various decision thresholds. The examples also illustrate that the probability of being the best reflects the confidence or strength of conviction of a classifier for its classification of the data. Both hypotheses

are supported by examples where only one test data set is used and by examples where performance is averaged over multiple test data sets.

Understanding the meaning of these performance measures is essential to the objective evaluation of competing classifiers. If classifiers are to be operated at a specific decision threshold, then classification accuracy may be a sufficient performance measure. However, if the choice of classifier for a given application is desired to be independent of a particular decision threshold, then AUC or the average metric distance from the diagonal should be used to differentiate between the classifiers. The average metric distance from the diagonal has the advantage of being a true metric which provides more insight about classifier differences when the ROC curves of competing classifiers overlap. Also, the average metric distance provides the evaluator with information about the stability of  $CA$  for changes in decision thresholds, which can be alternatively viewed as perturbations in the data. Finally, the probability of being the best can be used as a tie-breaker when the other performance measures for the competing classifiers are statistically or practically insignificant, or when the strength of conviction of a classifier for its classification is of primary concern.

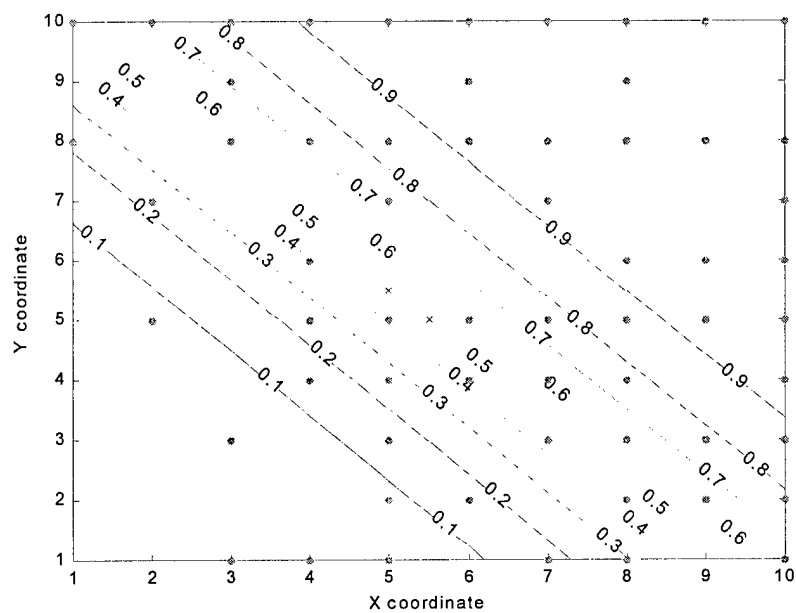


Figure 7.12 Average estimated Linear posterior probability contour plot for University of Wisconsin Breast Cancer Diagnosis Data Set (Dots represent actual Class 2 data points and X's mark class boundary).

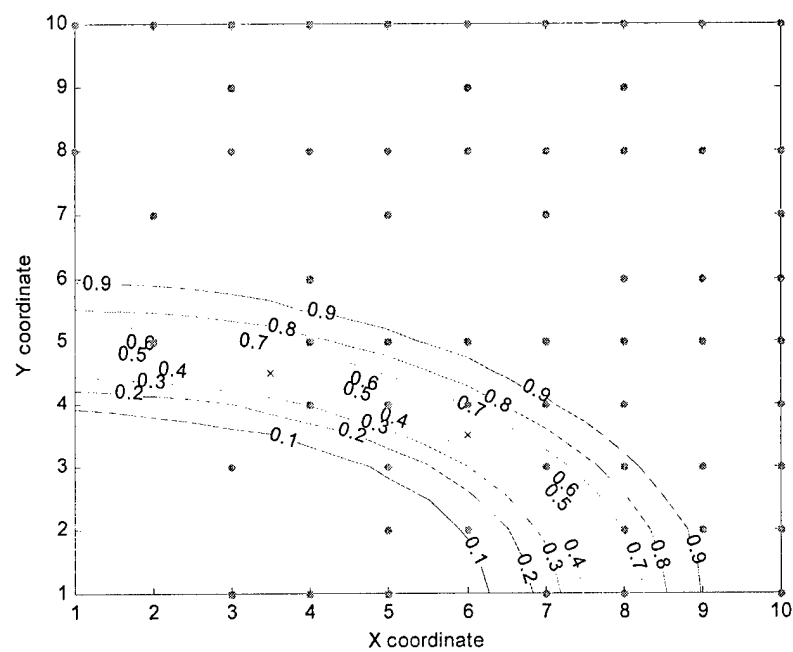


Figure 7.13 Average estimated Quadratic posterior probability contour plot for University of Wisconsin Breast Cancer Diagnosis Data Set (Dots represent actual Class 2 data points and X's mark class boundary).



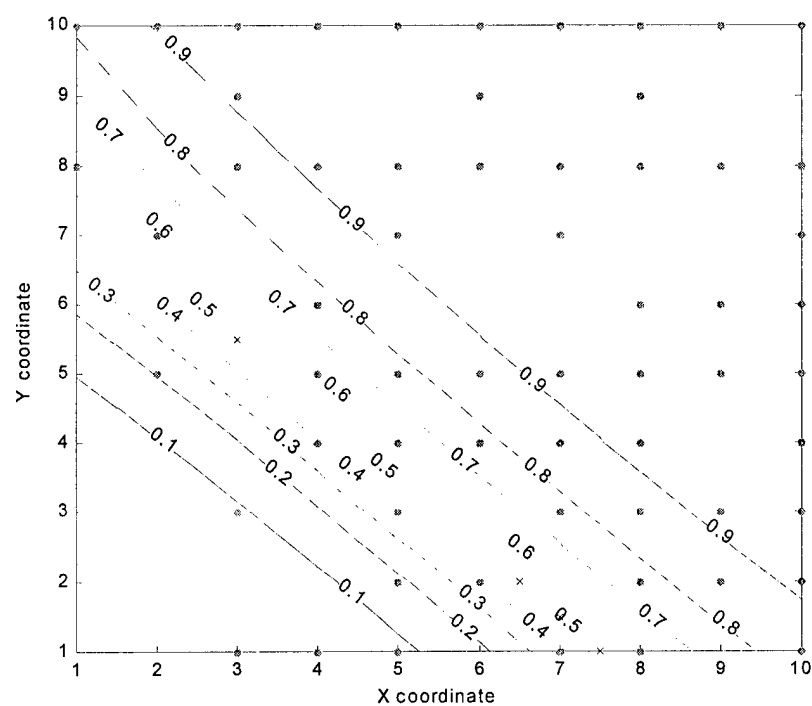


Figure 7.14 Average estimated MLP posterior probability contour plot for University of Wisconsin Breast Cancer Diagnosis Data Set (Dots represent actual Class 2 data points and X's mark class boundary).

## VIII. *Summary and Recommendations*

### 8.1 *Overview*

This dissertation research makes contributions towards the objective evaluation of competing classifiers. This chapter summarizes these contributions and also provides recommendations for future research.

### 8.2 *Contributions*

This section summarizes the contributions resulting from this research.

*8.2.1 Development of the Signal-to-Noise Ratio (SNR) Screening Method.* The initial stages of this research solidified ideas previously presented [33–35,63] on concepts for the Signal-to-Noise Ratio (SNR) screening method. These ideas were brought to practical fruition in an archival journal paper [14]. The SNR screening method uses the SNR saliency measure proposed in the paper to select a parsimonious set of salient features for an artificial neural network. Confidence in the SNR screening method and the SNR saliency measure are bolstered by comparisons to two other feature selection methods on three real-world problems.

*8.2.2 Background Reference on Performance Assessment and Performance Comparison Methods Used in Classifier Evaluation.* This dissertation provides a background reference on performance assessment and performance comparison methods used in classifier evaluation. The development of this background reference specifically for AFRL/SN on ATR has resulted in one technical report [2]. However, the performance assessment and performance comparison methods described in this background reference apply equally as well to a wide variety of other classification and detection problems.

*8.2.3 Proof of Convergence of Receiver Operating Characteristic (ROC) Curves.* This dissertation provides a proof of convergence of ROC curves. This ROC convergence theorem is

important because it provides the basis for a framework for the comparison of ROC curves and hence, the comparison of competing classifiers. The development of this proof has resulted in one submitted archival journal paper [7], one referee reviewed conference paper [5], and one technical report [6].

*8.2.4 Development of a New Methodology for Comparing ROC Curves.* This dissertation presents a family of metrics for comparing ROC curves. The development of these metrics as a useful tool in distinguishing between the ROC curves of competing classifiers resulted in one submitted archival journal paper [7], three referee reviewed conference papers [4, 5, 19], and two technical reports [3, 6]. This methodology is successfully applied to two extremely different applications, namely ATR and pilot workload detection.

*8.2.5 Development of a New Methodology Using a Multinomial Selection Procedure for Comparing Competing Classifiers.* This dissertation introduces a multinomial selection procedure as an alternative to ROC analysis for evaluating competing classifiers. This methodology is successfully applied to two extremely different applications, namely ATR and pilot workload detection. The results provide confidence in the multinomial selection procedure as a useful tool in distinguishing between competing classifiers.

### *8.3 Recommendations for Future Research*

There are many areas for future research and this section will list but just a few.

*8.3.1 Application of New Methodologies for Evaluating Competing Classifiers to Other Classifier Types and Other Problems.* This dissertation research uses three distinct classifier types:

1. Linear Statistical Classifier
2. Quadratic Statistical Classifier

### 3. Feed-forward MLP Classifier

More research in the application of the new methodologies presented in this dissertation for evaluating competing classifiers to other types of classifiers could be explored. Specifically, non-parametric classifiers such as those based on nearest neighbor methods [18, 60] and ANNs that allow for the encoding of time such as the Elman recurrent neural network (RNN) [18, 28] could be investigated. Also, the new methodologies could be applied to a variety of other classification problems, such as loan classification in business or galaxy classification in astronomy.

*8.3.2 Extension of New Methodology for Comparing ROC Curves to Multiple Probability Measures.* Future research may investigate the extension of the new methodology for comparing ROC curves developed in this dissertation to multiple probability measures. Instead of comparing standard two-dimensional (2-D) ROC curves, multi-dimensional ROC trajectories consisting of the usual probabilities of false alarm and detection along with a third, fourth, etc. probability measure could be compared. For example, a three-dimensional (3-D) ROC trajectory can be formed by adding the probability of rejection as the third dimension to the standard ROC curve [4]. The methodology developed in this dissertation for comparing ROC curves could be extended to compare these 3-D ROC trajectories.

*8.3.3 Development of a Systematic Methodology for Using Both Typical Performance Measures and Proposed Measures.* Future research may develop a systematic methodology for using both the typical performance measures, such as classification accuracy and area under the ROC curve along with the average metric distance from the diagonal line and the total probability of being the best classifier. Such a methodology would be valuable in the practical evaluation of competing classifiers.

*8.3.4 Development of a Hybrid Classifier Using the Classification Results of Competing Classifiers.* Another area of possible research is the development of better CSs by fusing the classifi-

cation results of the competing classifiers. Bayesian inference [36] could be applied to the estimated posterior probabilities generated by the classifiers for a specific class of interest. As a starting point, a simple fusion technique could be employed to rank order the classifiers according to the values of these posterior probabilities for each training data point. The conditional probability of the indication or in this simple technique, the order of the classifiers, given the specific class could then be computed. Bayes theorem could then be used to obtain the conditional probability of a specific class given the indication. Using an appropriate loss function, a new hybrid classifier could then be generated which predicts the class given the order of the individual classifiers for new data. More robust hybrid classifiers could be investigated by expanding the event space for the indication of the individual classifiers.

## Appendix A. Proof of ROC Convergence Theorem

### A.1 Overview

To prove the ROC Convergence Theorem the proof proceeds as follows:

1. Prove that the probabilities of false positive and true positive are consistent estimators.
2. Prove pointwise convergence for the estimated probability pair (false positive, true positive).
3. Prove that the integral of a real-valued random variable converges.
4. Prove that the sequence of ROC curves converges.

### A.2 Proof that the probabilities of false positive and true positive are consistent estimators

Let  $\Theta \subset \mathbb{R}$ . First, it will be shown that  $\hat{P}_{TP}^{(n)}(\omega, \theta)$  is a consistent estimator for  $P_{TP}(\theta)$  for each  $\theta \in \Theta$  and for each  $\omega \in \Omega$ . For  $\theta \in \Theta$  fixed and a given instantiation  $\omega \in \Omega$  of the data, define  $\hat{p}_n$  as

$$\hat{p}_n \equiv \hat{P}_{TP}^{(n)}(\omega, \theta) = \frac{\sum_{i=1}^n \chi_{[0, \theta]}(z_i | C_2)}{n} = \frac{\sum_{i=1}^n \eta_i}{n}. \quad (\text{A.1})$$

Note that  $\hat{p}_n$  is the sample mean of a random sample of size  $n$  of binary or Bernoulli random variables,  $\eta_i$ , so  $\hat{p}_n$  is also a random variable. Let  $S$  be the set from which feature vectors  $\mathbf{x}$  are drawn. Let  $\mathcal{D}^{(n)} \subset S$  be the set of feature vectors  $\mathbf{x}$  for finite  $n$ , i.e.,  $\mathcal{D}^{(n)} = \{\mathbf{x}_i \in \mathbb{R}^v : i = 1, \dots, 2n\}$  where  $v$  is the number of variables or features. Assume that  $\mathcal{D}^{(n)}$  converge to  $S$  in the Hausdorff metric,  $d_H$  [12], i.e., given  $\epsilon > 0$ , there exists  $N$  such that for all  $n > N$ ,

$$d_H(\mathcal{D}^{(n)}, S) < \epsilon \quad (\text{A.2})$$

then the Bernoulli random variables  $\eta_i$  have a mean  $\pi$  and positive variance  $\sigma^2$  for each  $i$

$$\mathbf{E}[\eta_i] = \pi \quad \text{and} \quad \mathbf{Var}[\eta_i] = \sigma^2 = \pi(1 - \pi) \quad (\text{A.3})$$

where  $\mathbf{E}$  denotes the Expectation operator  $\mathbf{Var}$  denotes the variance operator. Then the mean of  $\hat{p}_n$  is  $\pi$  since

$$\mathbf{E}[\hat{p}_n] = \mathbf{E}\left[\frac{1}{n} \sum_{i=1}^n \eta_i\right] = \frac{1}{n} \mathbf{E}\left[\sum_{i=1}^n \eta_i\right] = \frac{1}{n} \left[\sum_{i=1}^n \mathbf{E}[\eta_i]\right] = \frac{1}{n} \left[\sum_{i=1}^n \pi\right] = \frac{1}{n} n\pi = \pi \quad (\text{A.4})$$

and the variance of  $\hat{p}_n$  is  $\frac{\sigma^2}{n}$  since

$$\mathbf{Var}[\hat{p}_n] = \mathbf{Var}\left[\frac{1}{n} \sum_{i=1}^n \eta_i\right] = \frac{1}{n^2} \mathbf{Var}\left[\sum_{i=1}^n \eta_i\right] = \frac{1}{n^2} \left[\sum_{i=1}^n \mathbf{Var}[\eta_i]\right] = \frac{1}{n^2} \left[\sum_{i=1}^n \sigma^2\right] = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n} \quad (\text{A.5})$$

Let  $\varepsilon > 0$  and consider the probability

$$\Pr(|\hat{p}_n - \pi| \geq \varepsilon). \quad (\text{A.6})$$

Choose  $k > 0$  and choose  $N \geq (k \frac{\sigma}{\varepsilon})^2$ . Thus  $\varepsilon \geq k \frac{\sigma}{\sqrt{N}}$  and for every  $n > N$  we have  $\varepsilon > k \frac{\sigma}{\sqrt{n}}$ .

Using Chebyshev's Inequality

$$\Pr\left(|\hat{p}_n - \pi| \geq k \frac{\sigma}{\sqrt{n}}\right) \leq \frac{1}{k^2} \quad (\text{A.7})$$

which implies

$$\Pr(|\hat{p}_n - \pi| \geq \varepsilon) \leq \frac{1}{k^2} < \frac{\sigma^2}{\varepsilon^2 n} \quad (\text{A.8})$$

for every  $n > N$ . Therefore as  $n$  becomes larger

$$\lim_{n \rightarrow \infty} \Pr(|\hat{p}_n - \pi| \geq \varepsilon) \leq \lim_{n \rightarrow \infty} \frac{\sigma^2}{\varepsilon^2 n} = 0. \quad (\text{A.9})$$

Hence,  $\{\hat{p}_n\}$  converges to  $\pi$  in probability, which is denoted here as

$$\hat{p}_n \xrightarrow{P} \pi. \quad (\text{A.10})$$

Therefore,  $\hat{p}_n$  is a consistent estimator, which implies  $\hat{P}_{TP}^{(n)}(\omega, \theta)$  is a consistent estimator for all  $\theta \in \Theta$  and for all  $\omega \in \Omega$ . Hence, there exists  $P_{TP}(\theta)$  for each  $\theta \in \Theta$  and for each  $\omega \in \Omega$  such that  $\hat{P}_{TP}^{(n)}(\omega, \theta) \xrightarrow{P} P_{TP}(\theta)$ . Using the same technique above, one can prove similarly that  $\hat{P}_{FP}^{(n)}(\omega, \theta)$  is also a consistent estimator for all  $\theta \in \Theta$  and for all  $\omega \in \Omega$ , that is, there exists  $P_{FP}(\theta)$  for each  $\theta \in \Theta$  and for each  $\omega \in \Omega$  such that  $\hat{P}_{FP}^{(n)}(\omega, \theta) \xrightarrow{P} P_{FP}(\theta)$ . In summary then,  $\hat{P}_{TP}^{(n)}(\omega, \theta)$  and  $\hat{P}_{FP}^{(n)}(\omega, \theta)$  are consistent estimators for each  $\theta \in \Theta$  and for each  $\omega \in \Omega$ , that is

$$\Pr \left( \left\{ \hat{\mathbf{P}}^{(n)}(\omega, \theta) : \left| \hat{P}_{FP}^{(n)}(\omega, \theta) - P_{FP}(\theta) \right| \geq \varepsilon \right\} \right) \leq \varepsilon \quad \text{or} \quad \hat{P}_{FP}^{(n)}(\cdot, \theta) \xrightarrow{P} P_{FP}(\theta) \quad (\text{A.11})$$

$$\Pr \left( \left\{ \hat{\mathbf{P}}^{(n)}(\omega, \theta) : \left| \hat{P}_{TP}^{(n)}(\omega, \theta) - P_{TP}(\theta) \right| \geq \varepsilon \right\} \right) \leq \varepsilon \quad \text{or} \quad \hat{P}_{TP}^{(n)}(\cdot, \theta) \xrightarrow{P} P_{TP}(\theta). \quad (\text{A.12})$$

### A.3 Proof of pointwise convergence for the estimated probability pair

Since all metrics are equivalent (**Theorem III.1**), consider for now the Manhattan metric  $\rho_1$ . Therefore, given  $\varepsilon > 0$ ,  $\theta \in \Theta$ , and  $\omega \in \Omega$ , there exists  $N_{FP} > 0$  and  $N_{TP} > 0$ . Take  $N = \max\{N_{FP}, N_{TP}\}$  then for each  $n > N$

$$\begin{aligned} & \left\{ \hat{\mathbf{P}}^{(n)}(\omega, \theta) : \left| \hat{P}_{FP}^{(n)}(\omega, \theta) - P_{FP}(\theta) \right| + \left| \hat{P}_{TP}^{(n)}(\omega, \theta) - P_{TP}(\theta) \right| \geq \varepsilon \right\} \\ & \subset \left\{ \hat{\mathbf{P}}^{(n)}(\omega, \theta) : \left| \hat{P}_{FP}^{(n)}(\omega, \theta) - P_{FP}(\theta) \right| \geq \frac{\varepsilon}{2} \right\} \cup \left\{ \hat{\mathbf{P}}^{(n)}(\omega, \theta) : \left| \hat{P}_{TP}^{(n)}(\omega, \theta) - P_{TP}(\theta) \right| \geq \frac{\varepsilon}{2} \right\} \end{aligned} \quad (\text{A.13})$$



so,

$$\begin{aligned}
& \Pr \left( \left\{ \hat{\mathbf{P}}^{(n)}(\omega, \theta) : \rho_1 \left( \hat{\mathbf{P}}^{(n)}(\omega, \theta), \mathbf{P}(\theta) \right) \geq \varepsilon \right\} \right) \\
& \leq \Pr \left( \left\{ \hat{\mathbf{P}}^{(n)}(\omega, \theta) : \left| \hat{P}_{FP}^{(n)}(\omega, \theta) - P_{FP}(\theta) \right| \geq \frac{\varepsilon}{2} \right\} \cup \left\{ \hat{\mathbf{P}}^{(n)}(\omega, \theta) : \left| \hat{P}_{TP}^{(n)}(\omega, \theta) - P_{TP}(\theta) \right| \geq \frac{\varepsilon}{2} \right\} \right) \\
& \leq \Pr \left( \left\{ \hat{\mathbf{P}}^{(n)}(\omega, \theta) : \left| \hat{P}_{FP}^{(n)}(\omega, \theta) - P_{FP}(\theta) \right| \geq \frac{\varepsilon}{2} \right\} \right) + \Pr \left( \left\{ \hat{\mathbf{P}}^{(n)}(\omega, \theta) : \left| \hat{P}_{TP}^{(n)}(\omega, \theta) - P_{TP}(\theta) \right| \geq \frac{\varepsilon}{2} \right\} \right) \\
& \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.
\end{aligned} \tag{A.14}$$

Hence, the estimated probability pair  $\hat{\mathbf{P}}^{(n)}(\omega, \theta) = \left( \hat{P}_{FP}^{(n)}(\omega, \theta), \hat{P}_{TP}^{(n)}(\omega, \theta) \right)$  converges pointwise in probability to  $\mathbf{P}(\theta) = (P_{FP}(\theta), P_{TP}(\theta))$ , that is

$$\hat{\mathbf{P}}^{(n)}(\cdot, \theta) \xrightarrow{P} \mathbf{P}(\theta), \text{ for each } \theta \in \Theta. \tag{A.15}$$

Since all metrics are equivalent, convergence in Equation A.15 is valid with respect to every possible metric on  $\mathbb{R}^2$ .

#### A.4 Proof that the integral of a real-valued random variable converges

In this section a general result is first established. This result is then applied to the sequence of random ROC curves in the next section.

Let  $\mu$  be a measure on  $\mathfrak{R}$  (possibly Lebesgue) such that  $0 < \mu(\Theta) < \infty$ . Let  $\Theta \subset \mathfrak{R}$  be a set of decision thresholds. For each  $\theta \in \Theta$  and  $n \in \mathbb{N}$  let  $Y^{(n)}(\theta)$  be a real-valued random variable. Assume that  $Y^{(n)}(\theta) \xrightarrow{P} Y(\theta)$  as  $n \rightarrow \infty$  for each  $\theta \in \Theta$ , that is, given  $\varepsilon > 0$  and  $\theta \in \Theta$ , assume there exists a  $N = N(\varepsilon, \theta) \in \mathbb{N}$ , such that for all  $n > N$

$$\Pr \left( \left| Y^{(n)}(\theta) - Y(\theta) \right| \geq \varepsilon \right) < \varepsilon \tag{A.16}$$

or written more explicitly,

$$m\left(\left\{\omega \in \Omega : \left|Y^{(n)}(\omega, \theta) - Y(\omega, \theta)\right| \geq \varepsilon\right\}\right) < \varepsilon \quad (\text{A.17})$$

where  $m(S)$  is the probability measure of an event  $S$  and  $\Omega$  is the set of all possible random events under consideration. Assume for each  $n \in \mathbf{N}$  that  $Y^{(n)}(\omega, \cdot)$  is  $\mu$ -integrable over  $\Theta$ , almost everywhere  $\omega \in \Omega$ , that is,

$$\int_{\Theta} Y^{(n)}(\omega, \theta) d\mu(\theta) \text{ exists almost everywhere } \omega \in \Omega. \quad (\text{A.18})$$

**Theorem A.1.** Assume  $\Theta \subset \mathbb{R}$  such that  $0 < \mu(\Theta) < \infty$  and

1.  $Y^{(n)}(\omega, \cdot)$  is  $\mu$ -integrable on  $\Theta$  almost everywhere  $\omega \in \Omega$ .
2.  $\{Y^{(n)}\}$  is uniformly bounded, i.e., there exists  $B < \infty$  such that  $|Y^{(n)}(\omega, \theta)| \leq B$  for all  $n, \theta$  almost everywhere  $\omega \in \Omega$ .
3.  $Y^{(n)}(\cdot, \theta) \xrightarrow{p} Y(\cdot, \theta)$  for each  $\theta \in \Theta$

Then

$$\int_{\Theta} Y^{(n)}(\cdot, \theta) d\mu \xrightarrow{p} \int_{\Theta} Y(\cdot, \theta) d\mu$$

This theorem is proved below. First a useful Proposition is given.

**Proposition A.1.** Assume  $0 < \mu(\Theta) < \infty$  and  $Y^{(n)}(\cdot, \theta) \xrightarrow{p} Y(\cdot, \theta)$  for each  $\theta \in \Theta$ . If  $\varepsilon > 0$  and  $\delta > 0$ , there exists  $S = S(\varepsilon, \delta) \subset \Theta$  and  $N = N(\varepsilon, \delta) \in \mathbf{N}$  such that  $\mu(S) < \delta$  and

$$m\left(\left\{\omega \in \Omega : \left|Y^{(n)}(\omega, \theta) - Y(\omega, \theta)\right| \geq \varepsilon\right\}\right) < \varepsilon$$

for all  $n > N$  and  $\theta \in \Theta - S$ .

**Proof of Proposition A.1**

Let  $\varepsilon > 0$  and  $\delta > 0$ . For each  $n \in \mathbf{N}$  define

$$S_n \equiv \left\{ \theta \in \Theta : m \left( \left\{ \omega \in \Omega : \left| Y^{(k)}(\omega, \theta) - Y(\omega, \theta) \right| \geq \varepsilon \right\} \right) \geq \varepsilon \text{ for some } k \geq n \right\} \quad (\text{A.19})$$

**Claim 1.**  $S_n$  is a  $\mu$ -measurable set for each  $n \in \mathbf{N}$ .

**Proof of Claim 1.**

Since  $Y^{(n)}(\omega, \theta)$  and  $Y(\omega, \theta)$  are measurable functions on  $\Omega \times \Theta$  (with respect to  $m \times \mu$ ) then for  $\varepsilon > 0$  and every  $k \in \mathbf{N}$

$$T^{(k)} \equiv \left\{ (\omega, \theta) \in \Omega \times \Theta : \left| Y^{(k)}(\omega, \theta) - Y(\omega, \theta) \right| \geq \varepsilon \right\} \quad (\text{A.20})$$

is a  $m \times \mu$ -measurable set. By results in measure theory [15], the cross-section set

$$T_\theta^{(k)} = \left\{ \omega \in \Omega : \left| Y^{(k)}(\omega, \theta) - Y(\omega, \theta) \right| \geq \varepsilon \right\} \quad (\text{A.21})$$

is  $\mu$ -measurable for each  $\theta \in \Theta$  and  $M^{(k)}(\theta) \equiv m(T_\theta^{(k)})$  is a  $\mu$ -measurable function. Therefore,

$$\begin{aligned} W^{(k)} &\equiv \left\{ \theta \in \Theta : M^{(k)}(\theta) \geq \varepsilon \right\} \\ &= \left\{ \theta \in \Theta : m \left( \left\{ \omega \in \Omega : \left| Y^{(k)}(\omega, \theta) - Y(\omega, \theta) \right| \geq \varepsilon \right\} \right) \geq \varepsilon \right\} \end{aligned} \quad (\text{A.22})$$

is a  $\mu$ -measurable set. However  $S_n = \bigcup_{k \geq n} W^{(k)}$ , which implies  $S_n$  is a  $\mu$ -measurable set since a countable union of measurable sets is measurable [8].

**Claim 2.**  $S_n$  is decreasing, that is, if  $n_1 < n_2$ , then  $S_{n_1} \supset S_{n_2}$ .

**Proof of Claim 2.**

Let  $n_1 < n_2$ . Let  $\theta \in S_{n_2}$ , then

$$m\left(\left\{\omega \in \Omega : \left|Y^{(k)}(\omega, \theta) - Y(\omega, \theta)\right| \geq \varepsilon\right\}\right) \geq \varepsilon \quad (\text{A.23})$$

for some  $k \geq n_2$ , say  $k^*$ . Since  $n_2 > n_1$ , then  $k^* > n_1$  so the inequality is true for some  $k \geq n_1$ .

Thus,  $\theta \in S_{n_1}$ . Hence,  $S_{n_2} \subset S_{n_1}$ .

**Claim 3.**  $\bigcap_{n \in \mathbb{N}} S_n = \emptyset$ .

**Proof of Claim 3.**

Assume  $\bigcap_{n \in \mathbb{N}} S_n \neq \emptyset$ . Let  $\theta^* \in \bigcap_{n \in \mathbb{N}} S_n$ . Since  $Y^{(n)}(\cdot, \theta^*) \xrightarrow{p} Y(\cdot, \theta^*)$  then given  $\varepsilon > 0$  there exists  $N = N(\varepsilon, \theta^*) \in \mathbb{N}$  such that for all  $n > N$ ,  $m\left(\left\{\omega \in \Omega : \left|Y^{(n)}(\omega, \theta^*) - Y(\omega, \theta^*)\right| \geq \varepsilon\right\}\right) < \varepsilon$ . But  $\bigcap_{n \in \mathbb{N}} S_n \subset S_N$ , thus  $\theta^* \in S_N$ . Therefore,  $m\left(\left\{\omega \in \Omega : \left|Y^{(k)}(\omega, \theta^*) - Y(\omega, \theta^*)\right| \geq \varepsilon\right\}\right) \geq \varepsilon$  for some  $k > N$ . This is a contradiction. Therefore,  $\bigcap_{n \in \mathbb{N}} S_n = \emptyset$ .

Since  $S_1 \subset \Theta$ , then  $\mu(S_1) \leq \mu(\Theta) < \infty$  and  $\{\mu(S_n)\}_{n=1}^\infty \subset \mathbb{R}$  is a decreasing sequence bounded below by zero. Thus,  $\lim_{n \rightarrow \infty} \mu(S_n) = 0$ . There exists  $M = M(\delta) \in \mathbb{N}$  such that  $\mu(S_M) < \delta$ . For  $\theta \in \Theta - S_M$  then for all  $n > M$

$$m\left(\left\{\omega \in \Omega : \left|Y^{(n)}(\omega, \theta) - Y(\omega, \theta)\right| \geq \varepsilon\right\}\right) < \varepsilon \quad (\text{A.24})$$

thus,  $Y^{(n)}(\cdot, \theta) \xrightarrow{p} Y(\cdot, \theta)$  uniformly on  $\Theta - S_M$ . This completes the proof of Proposition A.1.

**Proof of Theorem A.1**

Let  $\varepsilon > 0$ . For each  $n \in \mathbb{N}$  and almost everywhere  $\omega \in \Omega$

$$\begin{aligned}
\left| \int_{\Theta} Y^{(n)}(\omega, \theta) d\mu - \int_{\Theta} Y(\omega, \theta) d\mu \right| &= \left| \int_{\Theta} [Y^{(n)}(\omega, \theta) - Y(\omega, \theta)] d\mu \right| \\
&\leq \int_{\Theta} |Y^{(n)}(\omega, \theta) - Y(\omega, \theta)| d\mu \\
&= \int_{\Theta} |Y^{(n)}(\omega, \theta) - Y(\omega, \theta)| d\mu + \int_{\Theta-S} |Y^{(n)}(\omega, \theta) - Y(\omega, \theta)| d\mu \\
&\leq \int_S |Y^{(n)}(\omega, \theta) + Y(\omega, \theta)| d\mu + \int_{\Theta-S} |Y^{(n)}(\omega, \theta) - Y(\omega, \theta)| d\mu \\
&\leq 2B\mu(S) + \int_{\Theta-S} |Y^{(n)}(\omega, \theta) - Y(\omega, \theta)| d\mu
\end{aligned} \tag{A.25}$$

Take  $L = L(\varepsilon, \delta) = \max \{N(\varepsilon, \delta), M(\delta)\}$  and  $S = S_L$  (as in Proposition A.1) then  $\mu(S) \leq \delta$  so that Equation A.25 becomes

$$\left| \int_{\Theta} Y^{(n)}(\omega, \theta) d\mu - \int_{\Theta} Y(\omega, \theta) d\mu \right| \leq 2B\delta + \int_{\Theta-S} |Y^{(n)}(\omega, \theta) - Y(\omega, \theta)| d\mu \tag{A.26}$$

This inequality implies

$$\begin{aligned}
&\left\{ \omega \in \Omega : \left| \int_{\Theta} Y^{(n)}(\omega, \theta) d\theta - \int_{\Theta} Y(\omega, \theta) d\theta \right| \geq \varepsilon \right\} \\
&\subset \left\{ \omega \in \Omega : 2B\delta + \int_{\Theta-S} |Y^{(n)}(\omega, \theta) - Y(\omega, \theta)| d\mu \geq \varepsilon \right\}
\end{aligned} \tag{A.27}$$

Take  $\delta = \frac{\varepsilon}{4B}$  so that

$$\begin{aligned}
&\left\{ \omega \in \Omega : \left| \int_{\Theta} Y^{(n)}(\omega, \theta) d\theta - \int_{\Theta} Y(\omega, \theta) d\theta \right| \geq \varepsilon \right\} \\
&= \left\{ \omega \in \Omega : \int_{\Theta-S} |Y^{(n)}(\omega, \theta) - Y(\omega, \theta)| d\mu \geq \frac{\varepsilon}{2} \right\}
\end{aligned} \tag{A.28}$$

Since  $Y^{(n)}(\omega, \cdot) - Y(\omega, \cdot)$  is bounded on  $\Theta$ , there exists an essential supremum  $M^{(n,\omega)}$  and an essential infimum  $m^{(n,\omega)}$  given by

$$\begin{aligned} M^{(n,\omega)} &= \operatorname{ess\,sup}_{\theta \in \Theta - S} \left| Y^{(n)}(\omega, \theta) - Y(\omega, \theta) \right| \\ m^{(n,\omega)} &= \operatorname{ess\,inf}_{\theta \in \Theta - S} \left| Y^{(n)}(\omega, \theta) - Y(\omega, \theta) \right| \end{aligned} \quad (\text{A.29})$$

By the Mean Value Theorem there exists  $C^{(n,\omega)} \in [m^{(n,\omega)}, M^{(n,\omega)}]$  such that

$$\int_{\Theta - S} \left| Y^{(n)}(\omega, \theta) - Y(\omega, \theta) \right| d\mu(\theta) = C^{(n,\omega)} \mu(\Theta - S) \quad (\text{A.30})$$

Therefore,

$$\int_{\Theta - S} \left| Y^{(n)}(\omega, \theta) - Y(\omega, \theta) \right| d\mu = C^{(n,\omega)} \mu(\Theta - S) \quad (\text{A.31})$$

$$\leq C^{(n,\omega)} \mu(\Theta)$$

$$\leq M^{(n,\omega)} \mu(\Theta)$$

$$= \operatorname{ess\,sup}_{\theta \in \Theta - S} \left| Y^{(n)}(\omega, \theta) - Y(\omega, \theta) \right| \mu(\Theta) \quad (\text{A.32})$$

Thus,

$$\begin{aligned} &\left\{ \omega \in \Omega : \int_{\Theta - S} \left| Y^{(n)}(\omega, \theta) - Y(\omega, \theta) \right| d\mu \geq \frac{\varepsilon}{2} \right\} \\ &\subseteq \left\{ \omega \in \Omega : \operatorname{ess\,sup}_{\theta \in \Theta - S} \left| Y^{(n)}(\omega, \theta) - Y(\omega, \theta) \right| \geq \frac{\varepsilon}{2\mu(\Theta)} \right\} \end{aligned} \quad (\text{A.33})$$

There exists  $\theta_{n,\omega} \in \Theta - S$  such that

$$M^{(n,\omega)} - \frac{\varepsilon}{4\mu(\Theta)} \leq \left| Y^{(n)}(\omega, \theta_{n,\omega}) - Y(\omega, \theta_{n,\omega}) \right| \leq M^{(n,\omega)} \quad (\text{A.34})$$

So,

$$M^{(n,\omega)} \leq \left| Y^{(n)}(\omega, \theta_{n,\omega}) - Y(\omega, \theta_{n,\omega}) \right| + \frac{\varepsilon}{4\mu(\Theta)} \quad (\text{A.35})$$

Therefore,

$$\begin{aligned} & \text{ess sup}_{\theta \in \Theta - S} \left| Y^{(n)}(\omega, \theta) - Y(\omega, \theta) \right| \mu(\Theta) \\ & \leq \left( \left| Y^{(n)}(\omega, \theta_{n,\omega}) - Y(\omega, \theta_{n,\omega}) \right| + \frac{\varepsilon}{4\mu(\Theta)} \right) \mu(\Theta) \\ & = \left| Y^{(n)}(\omega, \theta_{n,\omega}) - Y(\omega, \theta_{n,\omega}) \right| \mu(\Theta) + \frac{\varepsilon}{4} \end{aligned} \quad (\text{A.36})$$

So,

$$\begin{aligned} \frac{\varepsilon}{2} & \leq \int_{\Theta - S} \left| Y^{(n)}(\omega, \theta) - Y(\omega, \theta) \right| d\mu \\ & \leq \text{ess sup}_{\theta \in \Theta - S} \left| Y^{(n)}(\omega, \theta) - Y(\omega, \theta) \right| \mu(\Theta) \\ & \leq \left| Y^{(n)}(\omega, \theta_{n,\omega}) - Y(\omega, \theta_{n,\omega}) \right| \mu(\Theta) + \frac{\varepsilon}{4} \end{aligned} \quad (\text{A.37})$$

So,

$$\frac{\varepsilon}{4\mu(\Theta)} \leq \left| Y^{(n)}(\omega, \theta_{n,\omega}) - Y(\omega, \theta_{n,\omega}) \right| \quad (\text{A.38})$$

Therefore for any  $\theta_{n,\omega} \in \Theta - S$

$$\begin{aligned} & \left\{ \omega \in \Omega : \left| \int_{\Theta} Y^{(n)}(\omega, \theta) d\theta - \int_{\Theta} Y(\omega, \theta) d\theta \right| \geq \frac{\varepsilon}{2} \right\} \\ & \subset \left\{ \omega \in \Omega : \left| Y^{(n)}(\omega, \theta_{n,\omega}) - Y(\omega, \theta_{n,\omega}) \right| \geq \frac{\varepsilon}{4\mu(\Theta)} \right\} \end{aligned} \quad (\text{A.39})$$

Since  $Y^{(n)}(\omega, \cdot) \xrightarrow{P} Y(\omega, \cdot)$  uniformly on  $\Theta - S$ , then for all  $n \geq L$

$$\begin{aligned} & m \left( \left\{ \omega \in \Omega : \left| \int_{\Theta} Y^{(n)}(\omega, \theta) d\mu - \int_{\Theta} Y(\omega, \theta) d\mu \right| \geq \frac{\varepsilon}{2} \right\} \right) \\ & \leq m \left( \left\{ \omega \in \Omega : |Y^{(n)}(\omega, \theta_{n,\omega}) - Y(\omega, \theta_{n,\omega})| \geq \frac{\varepsilon}{4\mu(\Theta)} \right\} \right) < \frac{\varepsilon}{4\mu(\Theta)} \end{aligned} \quad (\text{A.40})$$

for all  $\theta_{n,\omega} \in \Theta - S$ . Therefore,

$$\int_{\Theta} Y^{(n)}(\cdot, \theta) d\mu \xrightarrow{P} \int_{\Theta} Y(\cdot, \theta) d\mu \text{ as } n \rightarrow \infty \quad (\text{A.41})$$

#### A.5 Proof of the convergence of the sequence of ROC curves

Let  $\rho$  be any metric on  $\mathbb{R}^2$  and  $\theta \in \Theta$ . Define  $Y^{(n)}(\omega, \theta) \equiv \rho(\hat{\mathbf{P}}^{(n)}(\omega, \theta), \mathbf{P}(\theta))$  then  $Y^{(n)}(\cdot, \theta)$  is a random variable. Since  $\hat{\mathbf{P}}^{(n)}(\cdot, \theta) \xrightarrow{P} \mathbf{P}(\theta)$ , for all  $\theta \in \Theta$  (Section A.3), then  $Y^{(n)}(\cdot, \theta) \xrightarrow{P} 0$  for each  $\theta \in \Theta$ . Choose  $\mu$  such that  $0 < \mu(\Theta) < \infty$  then

- $Y^{(n)}(\omega, \cdot)$  is  $\mu$ -integrable on  $\Theta$  almost everywhere  $\omega \in \Omega$  because  $Y^{(n)}(\omega, \cdot)$  is of bounded variation on  $\Theta$ .
- $\{Y^{(n)}\}$  is uniformly bounded, i.e.,  $|Y^{(n)}(\omega, \theta)| \leq 1$  for all  $n, \theta$  almost everywhere  $\omega \in \Omega$  since  $\hat{P}_{FP}^{(n)}(\omega, \theta) \leq 1$  and  $\hat{P}_{TP}^{(n)}(\omega, \theta) \leq 1$  which implies  $\rho(\hat{\mathbf{P}}^{(n)}(\omega, \theta), \mathbf{P}(\theta)) \leq 1$ .
- $Y^{(n)}(\cdot, \theta) \xrightarrow{P} 0$  for each  $\theta \in \Theta$ .

then according to Theorem A.1

$$\int_{\Theta} Y^{(n)}(\cdot, \theta) d\theta \xrightarrow{P} 0 \quad (\text{A.42})$$

Let  $Z^{(n)}(\theta) = g(Y^{(n)}(\theta))$  where  $g$  is any continuous function. Then Equation A.42 implies

$$\int_{\Theta} Z^{(n)}(\cdot, \theta) d\theta \xrightarrow{P} 0 \quad (\text{A.43})$$



Let  $V^{(n)}(\omega) = h \left( \int_{\Theta} Z^{(n)}(\cdot, \theta) d\theta \right)$  where  $h$  is any continuous function. Then Equation A.43 implies

$$V^{(n)}(\omega) \xrightarrow{P} 0. \quad (\text{A.44})$$

For the specific case where

$$g(x) = x^r \text{ where } r \in \mathbf{N} \quad (\text{A.45})$$

$$h(x) = g^{-1} = x^{\frac{1}{r}} \text{ where } r \in \mathbf{N}$$

the metric  $d_{\rho,r}(\hat{f}^{(n)}(\omega), f)$  can be written as

$$\begin{aligned} d_{\rho,r}(\hat{f}^{(n)}(\omega), f) &= \left( \int_{\Theta} \rho \left( \hat{\mathbf{P}}^{(n)}(\omega, \theta), \mathbf{P}(\theta) \right)^r d\theta \right)^{\frac{1}{r}} \\ &= \left( \int_{\Theta} Y^{(n)}(\omega, \theta)^r d\theta \right)^{\frac{1}{r}} \\ &= \left( \int_{\Theta} Z^{(n)}(\omega, \theta) d\theta \right)^{\frac{1}{r}} \\ &= V^{(n)}(\omega) \end{aligned} \quad (\text{A.46})$$

Therefore, Equation A.44 implies that given  $\varepsilon > 0$ , there exists  $N$  such that for all  $n > N$ ,

$$\Pr \left( \left\{ \omega \in \Omega : d_{\rho,r}(\hat{f}^{(n)}(\omega), f) \geq \varepsilon \right\} \right) < \varepsilon \quad (\text{A.47})$$

which is the result sought, namely that  $\{\hat{f}^{(n)}(\omega)\}$  converges to  $f$ , i.e., the sequence of estimated ROC functions  $\hat{f}^{(n)}(\omega)$  converge.

*Appendix B. Glossary of Acronyms and Abbreviations*

**AFB** Air Force Base

**AFOSR** Air Force Office of Scientific Research

**ANN** Artificial Neural Network

**ANNIE** Artificial Neural Networks in Engineering

**AFRL/HE** Air Force Research Laboratory Human Effectiveness Directorate

**AFRL/SN** Air Force Research Laboratory Sensors Directorate

**APC** Armored Personnel Carrier

**ATR** Automatic Target Recognizer/Recognition

**ATRWG** Automatic Target Recognition Working Group

**AUC** Area Under the ROC Curve

**AVC** All Vector Comparison

**BEM** Bechhofer, Elmaghraby, and Morse

**BLK** Blinks

**BTH** Breaths

**CA** Classification Accuracy

**CI** Confidence Interval

**CS** Classification System

**EEG** Electroencephalograph

**FFT** Fast Fourier Transform

**FOA** Focus of Attention

**FN** False Negative

**FP** False Positive

**HR** Heart Rate

**HRR** High Range Resolution

**IBI** Interbeat Interval

**IBLKI** Interblink Interval

**IBTHI** Interbreath Interval

**IFR** Instrument Flight Rules

**IX** Index

**JASA** Journal of the American Statistical Association

**LFC** Least Favorable Condition

**MBT** Main Battle Tank

**MLP** Multi-layer Perceptron

**MML** Mobile Missile Launcher

**MSE** Mean Square Error

**MSP** Multinomial Selection Problem/Procedure

**MSTAR** Moving and Stationary Target Acquisition and Recognition

**PCS** Probability of Correct Selection

**PEMS** Predict/Extract/Match/Search

**ROC** Receiver Operating Characteristic

**ROI** Region of Interest

**SAR** Synthetic Aperture Radar

**SNR** Signal-to-Noise Ratio

**SPG** Self-Propelled Gun

**SSE** Sum of Square Error

**T** Truck

**TN** True Negative

**TP** True Positive

**USAF** United States Air Force

**VFR** Visual Flight Rules

**W** Wilcoxon

## Bibliography

1. ALSING, S. G. An analysis of psychophysiological features for classifying pilot workload in crew aircraft using artificial neural networks. Tech. rep., Air Force Institute of Technology, Wright Patterson AFB OH, Feb 1998.
2. ALSING, S. G., AND BAUER, JR., K. W. Survey of statistical analysis and experimental design in ATR evaluation. Tech. Rep. WP98-05, Air Force Institute of Technology, Wright Patterson AFB OH, May 1998.
3. ALSING, S. G., AND BAUER, JR., K. W. Evaluation of competing classifiers for extremely unbalanced data using receiver operating characteristic type analysis and a multinomial selection procedure. Tech. Rep. WP99-01, Air Force Institute of Technology, Wright Patterson AFB OH, Jan 1999.
4. ALSING, S. G., BAUER, JR., K. W., AND BLASCH, E. P. Three-dimensional receiver operating characteristic trajectory concepts for the evaluation of target recognition algorithms faced with the unknown target detection problem. In *Proceedings of SPIE: Automatic Target Recognition IX* (Orlando FL, Apr 1999), vol. 3718, pp. 449-458.
5. ALSING, S. G., BAUER, JR., K. W., AND OXLEY, M. E. Convergence for receiver operating characteristic curves and the performance of neural networks. In *Intelligent Engineering Systems through Artificial Neural Networks* (St Louis MO, Nov 1999), vol. 9 of *Proceedings of Artificial Neural Networks in Engineering (ANNIE) International Conference*, pp. 947-952.
6. ALSING, S. G., BAUER, JR., K. W., AND OXLEY, M. E. Convergence for receiver operating characteristic curves. Tech. Rep. WP99-04, Air Force Institute of Technology, Wright Patterson AFB OH, Jul 1999.
7. ALSING, S. G., BAUER, JR., K. W., AND OXLEY, M. E. A family of metrics for comparing receiver operating characteristic curves. *submitted to Journal of the American Statistical Association* (Mar 2000).
8. APOSTLE, T. M. *Mathematical Analysis*. Addison-Wesley Publishing Company, Menlo Park, 1974.
9. AUTEN, J. G-LOC. Is the clue bag half full or half empty. *Torch* (Sep 1995), 8-11.
10. AUTOMATIC TARGET RECOGNIZER WORKING GROUP (ATRWG) NO. 86-001. *Target Recognizer Definitions and Performance Measures*. Wright Patterson AFB OH, 1986.
11. AUTOMATIC TARGET RECOGNIZER WORKING GROUP (ATRWG) NO. 88-006. *Application of Confidence Intervals to ATR Performance Evaluation*. Wright Patterson AFB OH, 1988.
12. BARNESLEY, M. *Fractals Everywhere*. Academic Press, New York, 1988.
13. BARTLE, R. G., AND SHERBERT, D. R. *Introduction to Real Analysis*. John Wiley & Sons, New York, 1992.
14. BAUER, JR., K. W., ALSING, S. G., AND GREENE, K. A. Feature screening using signal-to-noise ratios. *Neurocomputing* 31 (Mar 2000), 29-44.
15. BEAR, H. S. *A Primer of Lebesgue Integration*. Academic Press, San Diego, 1995.
16. BECHHOFFER, R. E., SANTNER, T. J., AND GOLDSMAN, D. M. *Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons*. John Wiley & Sons, New York, 1995.

17. BELUE, L. M. *Selecting Optimal Experiments for Multiple Multilayer Perceptrons*. Ph.D. dissertation, Air Force Institute of Technology, Wright-Patterson AFB, OH, March 1995.
18. BISHOP, C. M. *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, 1995.
19. BLASCH, E. P., ALSING, S. G., AND BAUER, JR., K. W. Comparison of bootstrap and prior-probability synthetic data balancing methods for SAR target recognition. In *Proceedings of SPIE: Algorithms for Synthetic Aperture Radar Imagery VI* (Orlando FL, Apr 1999), vol. 3721, pp. 740-747.
20. BRADLEY, A. P. The use of area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* 30 (1997), 1145-1159.
21. CATLIN, A. E., BAUER, JR., K. W., MYKYTKA, E. F., AND MONTGOMERY, D. C. System comparison procedures for automatic target recognition systems. *Naval Research Logistics* 46 (1999), 357-371.
22. CHOW, C. K. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory* IT-16 (1970), 41-46.
23. DEMUTH, H., AND BEALE, M. *Neural Network Toolbox User's Guide, Version 3.0*. The Mathworks Inc, Natick, 1998.
24. DILLON, W. R., AND GOLDSTEIN, M. *Multivariate Analysis*. John Wiley & Sons, New York, 1984.
25. DORFMAN, D. D., AND ALF, JR., E. Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals-rating-method data. *Journal of Mathematical Psychology* 6 (1969), 487-496.
26. EAST, J. A. Feature selection for predicting pilot mental workload. MS thesis, Air Force Institute of Technology, Wright-Patterson AFB, OH, Mar 2000.
27. EGAN, J. P. *Signal Detection Theory and ROC Analysis*. Academic Press, New York, 1975.
28. ELMAN, J. L. Finding structure in time. *Cognitive Science* 14 (1972), 179-211.
29. FLEISS, J. L. *Statistical Methods for Rates and Proportions*. John Wiley & Sons, New York, 1981.
30. FOGEL, D. B., WASSON, E. C., BOUGHTON, E. M., PORTO, V. W., AND ANGELINE, P. J. Linear and neural models for classifying breast masses. *IEEE Transactions on Medical Imaging* 17 (Jun 1998), 485-488.
31. FOLEY, D. H. Considerations of sample and feature size. *IEEE Transactions on Information Theory* 18 (1972), 618-626.
32. GREEN, D. M., AND SWETS, J. A. *Signal Detection Theory and Psychophysics*. John Wiley & Sons, New York, 1966.
33. GREENE, K. A., BAUER, JR., K. W., KABRISKY, M., ROGERS, S. K., RUSSELL, C. A., AND WILSON, G. F. A preliminary investigation of selection of EEG and psychophysiological features for classifying pilot workload. In *Intelligent Engineering Systems through Artificial Neural Networks* (St Louis MO, Nov 1996), vol. 6 of *Proceedings of Artificial Neural Networks in Engineering (ANNIE) International Conference*, pp. 691-697.
34. GREENE, K. A., BAUER, JR., K. W., KABRISKY, M., ROGERS, S. K., AND WILSON, G. F. Estimating pilot workload using Elman recurrent networks: A preliminary investigation. In *Intelligent Engineering Systems through Artificial Neural Networks* (St Louis MO, Nov 1997),

- vol. 7 of *Proceedings of Artificial Neural Networks in Engineering (ANNIE) International Conference*, pp. 703-708.
35. GREENE, K. A., BAUER, JR., K. W., KABRISKY, M., ROGERS, S. K., AND WILSON, G. F. Determining the memory capacity of an Elman recurrent neural network. In *Intelligent Engineering Systems through Artificial Neural Networks* (St Louis MO, Nov 1998), vol. 8 of *Proceedings of Artificial Neural Networks in Engineering (ANNIE) International Conference*, pp. 37-42.
  36. HALL, D. L. *Mathematical Techniques in Multisensor Data Fusion*. Artech House, Boston, 1992.
  37. HAN, R. Y., AND CLARK, R. J. Characterization and evaluation of automatic target recognizer performance. In *Proceedings of the International Society for Optical Engineering* (Bellingham, 1984), vol. 504, SPIE, pp. 341-351.
  38. HARRUP, G. K. ROC analysis of IR segmentation techniques. MS thesis, Air Force Institute of Technology, Wright-Patterson AFB, Ohio, Dec 1994.
  39. HIGHLEYMAN, W. H. The design and analysis of pattern recognition experiments. *Bell System Technical Journal* 41 (1962), 723-744.
  40. HILDEBOLT, C. F., VANNIER, M. W., SHROUT, M. K., AND PILGRAM, T. K. ROC analysis of observer-response subjective rating data - application to periodontal radiograph assessment. *American Journal of Physical Anthropology* 84 (1991), 351-361.
  41. IRVING, W. W. A system behavioral model: Take 2. Tech. rep., ALPHATECH, Burlington, MA, Oct 1997.
  42. IRVING, W. W., AND WASHBURN, R. B. Performance model for MSTAR system. Tech. Rep. TM-473, ALPHATECH, Burlington, MA, Feb 1996.
  43. IRVING, W. W., AND WISSINGER, J. Procedures for ROC curve generation with the proposed drop 5 MSTAR algorithm. Tech. rep., ALPHATECH, Burlington, MA, Oct 1997.
  44. JACHIMCZYK, W. R. Enhancements of Pose-tagged Partial Evidence Fusion for Automatic Target Recognition. MS thesis, Worcester Polytechnic Institute, Worcester, MA, May 1997.
  45. JAMES, M. *Classification Algorithms*. John Wiley & Sons, New York, 1985.
  46. LAW, A. M., AND KELTON, W. D. *Simulation Modeling and Analysis*, second ed. McGraw-Hill, Inc., New York, 1991.
  47. LLOYD, C. J. Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. *Journal of the American Statistical Association* 93, 444 (Dec 1998), 1356-1364.
  48. MAHALANOBIS, P., AND MAHALANOBIS, A. Statistical inference for automatic target recognition systems. *Applied Optics* 33 (1994), 6823-6825.
  49. MCNEIL, B. J., AND HANLEY, J. A. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143 (1982), 29-36.
  50. MCNEIL, B. J., AND HANLEY, J. A. Statistical approaches to the analysis of receiver operating (ROC) curves. *Medical Decision Making* 4 (1984), 137-150.
  51. METZ, C. E. ROC methodology in radiologic imaging. *Investigative Radiology* 21 (1986), 720-733.

52. METZ, C. E., WANG, P., AND KRONMAN, H. B. A new approach for testing the significance of differences between ROC curves measured from correlated data. In *Information Processing in Medical Imaging VIII* (Hague, 1984), F. Deconick, Ed., Martinus Nijhof, pp. 432-435.
53. MILLER, J. O., NELSON, B. L., AND REILLY, C. H. Efficient Multinomial selection in simulation. *Naval Research Logistics* 45 (1998), 459-482.
54. MITCHELL, R. A., AND WESTERKAMP, J. J. A statistical feature based classifier for robust high range resolution radar target identification. Tech. rep., Air Force Research Laboratory Sensors Directorate (AFRL/SN), Wright-Patterson AFB OH, 1998.
55. MONTGOMERY, D. C. *Design and Analysis of Experiments*, fourth ed. John Wiley & Sons, New York, 1997.
56. NAYLOR, A. W., AND SELL, G. R. *Linear Operator Theory in Engineering and Science*. Springer-Verlag, New York, 1982.
57. NETRELLA, M. G. *Experimental Statistics*. Department of Commerce, Washington DC: Government Printing Office, 1996.
58. PETERSON, W. W., BIRDSALL, T. G., AND FOX, W. C. The theory of signal detectability. *Transactions of the IRE Professional Group on Information Theory PGIT-4* (1954), 171-212.
59. RICHARD, M. D., AND LIPPMANN, R. P. Neural network classifiers estimate Bayesian a posteriori probabilities. *Neural Computation* 3 (1991), 461-483.
60. RIPLEY, B. D. *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, 1996.
61. ROSS, T. D., WESTERKAMP, L. A., ZELINO, E. G., AND BURNS, T. J. Extensibility and other model-based ATR evaluation concepts. In *Proceedings of the International Society for Optical Engineering (SPIE)* (Orlando FL, Apr 1997), vol. 3070, pp. 213-222.
62. RUCK, D. W., ROGERS, S. K., KABRISKY, M., OXLEY, M. E., AND SUTER, B. W. The multilayer perceptron as an approximation to a Bayes optimal discriminant function. *IEEE Transactions on Neural Networks* 1, 4 (1990), 296-298.
63. SUMRELL, D. B. *An investigation of preliminary feature screening using signal-to-noise ratios*. Ph.D. dissertation, Air Force Institute of Technology, Wright-Patterson AFB, OH, 1996.
64. SWETS, J. A., AND PICKETT, R. M. *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*. Academic Press, New York, 1982.
65. SWINGLER, K. *Applying Neural Networks, A Practical Guide*. Academic Press, London, 1996.
66. UNIVERSITY OF CALIFORNIA-IRVINE (UCI). *Machine Learning Repository*. WWWeb, <http://www.ics.uci.edu/mlearn/>, 1999.
67. VELTEN, V., ROSS, T., MOSSING, J., WORRELL, S., AND BRYANT, M. Standard SAR ATR evaluation experiments using the MSTAR public release data set. Tech. rep., Air Force Research Laboratory Sensors Directorate (AFRL/SN), Wright-Patterson AFB OH, 1998.
68. WALD, A. *Sequential Analysis*. John Wiley & Sons, New York, 1947.



### *Vita*

Lieutenant Colonel Stephen G. Alsing was born 31 January 1961 in Suffren, New York. He graduated from Waldwick High School (New Jersey) as the Valedictorian in 1979 and attended Michigan State University in East Lansing, Michigan. In 1983, Lieutenant Colonel Alsing graduated from Michigan State with a Bachelor of Science Degree in Astrophysics and was commissioned as a distinguished graduate of the Air Force Reserve Officer Training Corps. In 1984, he graduated from Michigan State with a Master of Science Degree in Physics. His first assignment was to Reese AFB, Texas to attend Undergraduate Pilot Training (UPT). Upon graduating UPT in 1985 as the Academic Ace (number one in academics), Lieutenant Colonel Alsing served as a pilot, aircraft commander, instructor pilot, evaluator pilot, and flight commander in the KC-135A Stratotanker at Barksdale Air Force Base in Louisiana and at Minot Air Force Base in North Dakota. In 1993, he returned to academia as an Assistant Professor of Physics and as the Observatory Director at the United States Air Force Academy in Colorado. Lieutenant Colonel Alsing entered the School of Engineering, Air Force Institute of Technology (AFIT) in August 1996 to pursue a Ph.D. in Operations Research.

REPORT DOCUMENTATION PAGE				Form Approved OMB No. 0704-0188	
<p>The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.</p> <p><b>PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.</b></p>					
1. REPORT DATE (DD-MM-YYYY) 06-03-2000		2. REPORT TYPE Ph.D. Dissertation		3. DATES COVERED (From - To) Jan 1999 - Mar 2000	
4. TITLE AND SUBTITLE THE EVALUATION OF COMPETING CLASSIFIERS				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S) Stephen G. Alsing, Lieutenant Colonel, USAF				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Air Force Institute of Technology (AFIT) Wright-Patterson AFB OH 45433-7765				8. PERFORMING ORGANIZATION REPORT NUMBER AFIT/DS/ENS/00-01	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Capt Erik P. Blasch, Ph.D. AFRL/SNAT 2241 Avionics Cir Wright-Patterson AFB OH 45433-7321 (937)-255-8639x3305 (DSN 785)				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution unlimited					
13. SUPPLEMENTARY NOTES Advisor: Dr. Kenneth W. Bauer, Jr., Professor, Department of Operational Sciences, Air Force Institute of Technology Email: Kenneth.Bauer@afit.af.mil					
14. ABSTRACT The purpose of this dissertation research is to advance the knowledge of classifier evaluation. The basis of the research is a commonly used evaluation tool in automatic target recognition (ATR) and medical applications, called the receiver operating characteristic (ROC) curve. A proof of convergence with respect to increasing sample size for these ROC curves is provided. This ROC convergence theorem is important because it provides the basis for a framework for the comparison of ROC curves and hence, the comparison of classifiers. A demonstration is given to show how this framework can be employed using metrics that provide more insight about classifier differences than the typical area under the curve performance index used in ROC analysis. As an alternative to ROC type analyses, a method for using a multinomial selection procedure to evaluate competing classifiers is presented and demonstrated. A comparison is then made between the methodologies introduced in this research and typical approaches. Both ATR and pilot workload applications are used to make these comparisons. A review of the interpretations of the typical performance measures used is given along with interpretations for the proposed performance measures introduced in this dissertation. Finally, research contributions are summarized and future directions highlighted.					
15. SUBJECT TERMS Automatic target recognition, classifier, convergence, multinomial selection procedure, pilot workload, receiver operating characteristic.					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES 162	19a. NAME OF RESPONSIBLE PERSON Dr. Kenneth W. Bauer, Jr.
a. REPORT U	b. ABSTRACT U	c. THIS PAGE U			19b. TELEPHONE NUMBER (Include area code) (937) 255-6565 x4328 (DSN 785)